Making Email Actionable: The Identification and Use of Obligation Acts in Workplace Email

By

Andrew Lampert

This thesis is presented for the degree of
Doctor of Philosophy
at Macquarie University
Department of Computing
July 29th 2013

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, the original work of the author. All verbatim extracts have been distinguished by quotations, and all sources of information have been specifically acknowledged. This work has not been submitted for a degree or any other qualification to any other university or institution.

Andrew Lampert

Abstract

Email is a key communication medium in business environments, where it is often used to assign and delegate tasks. Existing research has established that task-oriented communication is built upon the exchange of request and commitment speech acts—collectively, *obligation acts*—between interlocutors, but email software has so far ignored this insight; it has not adapted to support task management, despite its popularity as a medium for such workflows. The lack of task awareness in email software has been repeatedly highlighted as a key factor in the 'information overload' that burdens many email users. In particular, the difficulty of distilling tasks from the ever-increasing email flow leads to obligations that remain unfulfilled.

This thesis explores how to address this problem by making email more actionable. We begin by analysing data from a series of annotation experiments through which we gathered independent human judgements about requests and commitments across a collection of more than 2000 real-world email messages. These annotated messages provide insight into how obligation acts are realised and interpreted. We identify and analyse a range of complex phenomena involved in these speech acts, and provide definitions for identifying them in email.

Building on this analysis, we then present effective computational techniques for detecting obligation acts at three levels of granularity within email messages: the message, paragraph and sentence levels. Message-level identification determines whether or not an email message contains obligation acts, and aims to assist users to triage their messages by focusing on those containing actionable content. Paragraph-level identification builds on this to classify each paragraph in the same manner; this enables, for example, the production of extractive summaries of messages. Finally, sentence-level identification classifies each sentence in an email message, and allows requests and commitments to be extracted to external task lists. We use our annotated email data to train supervised machine learning algorithms for each of these classification tasks. These classifiers also exploit a novel classification system that segments the text of email messages into different functional zones, identifying material such as signatures, advertising, and quoted reply content. This enables our obligation classifiers to focus on only relevant email text when identifying requests and commitments.

In sum, this thesis provides both theoretical and practical foundations for managing obligation acts in real email data. Our empirically-grounded analysis and the software tools developed on the basis of this analysis demonstrate how it is possible to make email actionable today, as well as providing a platform for future task-related email research.

To Michelle, with my deepest thanks for your unconditional love, your patience, and your selfless support.

Dance with me
On a cliff-top tee
Until our feet are sore.

Laugh with me
Above the bubbling sea
Until our spirits soar.

Walk with me
For eternity
On love's cushioned floor.

Acknowledgements

The road to completing a thesis is never a one-person journey. Like Frodo Baggins on his great quest, I am only at my journey's end because I have been supported, both professionally and personally, by a Fellowship of amazing and inspiring people.

Firstly, thank you to my wife, Michelle. Quite simply, Shell, I could never have completed this journey without your unwavering support and encouragement. Our three beautiful children, Ronan, Finn and Greta all have the dubious honour of being 'thesis kids', born in that seemingly never-ending period between the ACL conference in Sydney in July 2006, and submission in July 2013. Michelle, for all that you are and all that you do to nurture my dreams and aspirations, I love you and thank you.

I am, of course, deeply indebted to my two supervisors, friends, mentors and guiding academic stars, Dr. Cécile Paris and Professor Robert Dale. Somehow you never (quite) lost faith in my ability to actually slay this thesis dragon. Thank you for your encouragement, discussion, corrections, scribbled notes, ideas, constructive criticism and for our many hastily convened Skype meetings at airports and cities around the world. You both challenge me in just the right ways to have made this thesis something it could never have otherwise been. I could not hope for better academic 'parents'.

I also extend my sincere thanks to my thesis examiners, Professor Candace Sidner, Associate Professor Giuseppe Carenini and Dr. Lawrence Cavedon. Reviewing a thesis is a demanding task, and I am grateful for your thoughtful comments and encouraging feedback.

Ronan, Finn and Greta, none of you have yet known life without Daddy's PhD. Yay, for now having more time to spend together as a family! If you ever look back on this thesis, I hope you dwell not on the many hours in which Daddy's attention was otherwise occupied, but instead on the fact that even seemingly insurmountable tasks *can* in fact be conquered with the right amount of persistence and support. Take this lesson with you in life!

To Brenda, Robert, Cécile, Michelle and Liz I am deeply grateful for your countless hours of painstaking email annotation and discussion of annotation disagreements! None of the work in this thesis would have been possible without your input.

A big thank you to Dr. Ross Wilkinson, Dr. Cécile Paris, Dr. Peter Bailey, Dr. Mark Cameron, Dr. Alex Zelinsky and the CSIRO ICT Centre for supporting my PhD while I was a member of staff. Thank you also to all my colleagues and friends from E6B, Marsfield and ANU, and especially to my fellow i2tech staff and 'Cécile's troops'. From the ANU side, a particular shout out to Daniel Breese for his help with the Outlook plug-in.

I owe a debt of gratitude to CSIRO, Macquarie University and HCSNet (the Australian Research Council Network in Human Communication Science) for grants that allowed me to travel to Chicago, Bowral, Singapore, Blackheath, Melbourne and Los Angeles so I could present and discuss my research with national and international peers. Thank you also to the Australasian Language Technology Association (ALTA), for consistently providing a welcoming community for local NLP and CL researchers.

To my now widely-flung comrades in the Markists group, thank you for instilling a sense of camaraderie and fun into the PhD process. To Associate Professor Mark Dras (the benevolent dictator), Dr. Stephen Wan (thanks for being such an amazing friend over so many years, so smart and thoughtful, and the best office-mate one could hope for!), Dr. Jean-Philippe Prost, Dr. Jette Viethen, Dr. Simon Zwarts, Dr. Mary Gardiner, Dr. Luiz Pizzato, Dr. Pawel Mazur, Marc Tilbrook, Dr. Elena Akhmatova, Jojo Wong, Teresa Lynn, Suzy Howlett, Yasaman Motazedi and Chris Rauchle: thanks for your friendship; you were the heart of the distinctive Macquarie CLT group culture that is at once inclusive, rigorous and supportive. Thanks for being there to listen, drink, teach, console, question and debate. To Associate Professor Steve Cassidy, Professor Mark Johnson, Dr. Diego Mollá Aliod, Dr. Rolf Schwitter, Dr. François Lareau and the rest of my colleagues at the CLT, thank you for your questions, conversation and contributions, both big and small.

And last, but certainly not least, to my parents, Wendy and Nigel, and my sister, Elisabeth, who learned ever more subtle and indirect ways to ask about my thesis progress in the face of my frequent unwillingness to talk about it. Thanks for your part in moulding me into the person I am today.

Publications Arising from this Thesis

Below is a list of peer-reviewed publications containing work from this thesis:

- Andrew Lampert, Robert Dale and Cécile Paris *Detecting Emails Containing Requests for Action.* In Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 984–992, Los Angeles, USA, June 1–6 (2010)

- Andrew Lampert, Robert Dale and Cécile Paris *Segmenting Email Message Text into Zones.* Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pages 919–928, Singapore, August 6–7 (2009)

- Andrew Lampert, Hong-Linh Truong, Simon Scerri and Michal Laclavik *1st International Workshop on E-mails in E-commerce and Enterprise* In Proceedings of 11th IEEE Conference on Commerce and Enterprise Computing (CEC), pages xxiii–xxiv, Vienna, Austria, July 20 (2009)

- Andrew Lampert *Email in the Australian National Corpus.* In Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, pages 55–60 (2009)

- Andrew Lampert, Robert Dale and Cécile Paris *Requests and Commitments in Email are More Complex Than You Think: Eight Reasons to be Cautious.* In Proceedings of Australasian Language Technology Workshop (ALTA), pages 55–63, Hobart, Australia, December 8–10 (2008)

- Andrew Lampert, Robert Dale and Cécile Paris *The Nature of Requests and Commitments in Email Messages.* In Proceedings of EMAIL-08: the AAAI Workshop on Enhanced Messaging, pages 42–47, Chicago, USA, July 13 (2008)

- Andrew Lampert, Cécile Paris and Robert Dale *Can Requests-for-Action and Commitments-to-Act be Reliably Identified in Email Messages?.* In Proceedings of the 12th Australasian Document Computing Symposium (ADCS), pages 48–55, Melbourne, Australia, December 10 (2007) **Best Presentation Award**

- Andrew Lampert, Robert Dale and Cécile Paris *Classifying Speech Acts using Verbal Response Modes.* In Proceedings of the 2006 Australasian Language Technology Workshop (ALTW), pages 34–41, Sydney, Australia, November 30–December 1 (2006) **Best Paper Award**

Contents

List of Figures

List of Tables

"[Managers] would like to be able to track outstanding promises they have made to others, promises made to them, requests they've made that have not been met and requests made of them that they have not fulfilled."

Denise E Murray (1991, p. 27)

1 Introduction

Email is one of the most successful software applications yet devised, with many hundreds of millions of people around the world relying on it daily for personal and professional communications. Although recent trends, including the use of social messaging platforms such as Facebook and Twitter, are threatening to break its dominance for personal communication, email remains the *de facto* standard for online communication, particularly in the workplace.

As part of its role in business communication, email is widely used for assigning and delegating tasks. Unfortunately, despite its popularity as a medium for working with tasks, email software generally does not provide well integrated support for task management. A prime example of this is that email users are not easily able to distinguish and manage task-related content separately from other text in their email messages. The lack of general task awareness in email software has been repeatedly highlighted as a key factor in the 'information overload' that burdens many users.

Existing research has established that task-oriented communication is built upon the exchange of DIRECTIVE speech acts, which request or demand action from the receiver, and COMMISSIVE speech acts, which commit the author or speaker to action. Throughout the thesis, we refer to directive utterances which place an obligation on one or more recipients as REQUESTS, and commissive utterances which place an obligation on the sender (or any non-recipient) as COMMITMENTS. Collectively, we refer to request and commitment acts as OBLIGATION ACTS. Requests and commitments form the basis of task-oriented communication, but email software has so far ignored this insight.

To fill this gap, this thesis provides a comprehensive exploration of how people issue requests and make commitments in real-world, workplace email communication. We build on the insights from our analysis to design, develop and evaluate a series of automated systems designed to improve how people work with task-related content in their email communications. Our goal is to provide better support for users to manage actionable content in email.

We begin this chapter with an overview of the history of email, and the technology and standards that define the content and transport of email messages. This provides a grounding for our review of the problem of email overload experienced by the many users who struggle to cope with their overflowing inboxes. We then look at how email has been adopted in the workplace, and in particular, how it is commonly applied to managing tasks in the workplace, the focus of this thesis.

We close the chapter by outlining the structure and contributions of this thesis. We begin with our theoretical and empirical analysis of how people make requests and commitments in email. Our work builds on influential ideas proposed by Winograd and Flores (1986) in taking a 'language/action' perspective and identifying speech acts in email. This approach differs from that of most existing email systems, which have for the most part treated the content of email messages as homogeneous bags of words. We then describe our work designing, building and evaluating systems to automate the detection of obligation acts at a range of different granularities within workplace email messages. Finally we introduce the integration of our systems into a commercial email software suite, with a view to creating a platform for conducting future email research in real-world contexts.

1.1 A Brief History of Email

The first networked email message was sent in late 1971 by BBN engineer Ray Tomlinson (Tomlinson, 2006). Local messaging programs that allowed users to exchange messages from different accounts on the same computer had existed for some time, since at least the early 1960s. It was Tomlinson, however, who extended the existing SNDMSG application for the Digital PDP-10 computer to allow messages to be sent to users at other computers connected over ARPANET, the precursor to the modern Internet. As part of this modest technological change, Tomlinson also introduced the now ubiquitous '@' symbol to distinguish recipients on remote machines from local recipients.

Tomlinson was at first reluctant to trumpet his innovation; his colleague at the time, Jerry Burchfiel, told Forbes magazine in 1998: "When he (Tomlinson) showed it to me, he said, 'Don't tell anyone! This isn't what we're supposed to be working on.'" (Cavender, 1998).[1] It was not long, however, before the appeal of fast, cheap, convenient communication over short or long distances began to win over the first of what would become billions of email users.[2] Larry Roberts, a director of DARPA, the government agency that ran the Advanced Research Projects Agency Network (ARPANET) on which email was transported, was an early user of Tomlinson's modified SNDMSG system and began using electronic mail for almost all his communication needs. That, in turn, encouraged researchers dependent on DARPA for research funding to also

[1] Tomlinson politely disputes this version of events; his take is that it was a colleague who made this comment, and that it was actually "said in jest because we were, after all, investigating ways in which to use the ARPANET"(Tomlinson, 2006).

[2] The Radicati Group estimates that, in 2011, there were 3.1 billion active email accounts; see http://www.radicati.com/?p=7269.

start using email for communication. At this point, the network effect had begun, and email quickly went from being a convenience to becoming the 'killer application' of the Internet.[3]

More than 40 years after Tomlinson's first message was sent, email has become an integral part of many people's lives, both within and outside the workplace. Hundreds of millions of people around the world rely on email for personal and professional communication. In a comprehensive survey from the United States, 86% of those who used email at work stated that it was essential or important to their work (Fallows, 2002). Its use also continues to grow: in research comparing email use between 1996 and 2006, Fisher et al. (2006) found that the average number of emails stored in an email account had increased tenfold from 2,482 messages in 1996 to 28,660 in 2006.

Many studies of email communication in the workplace have highlighted the diverse uses of email. In their survey of related research, Whittaker, Bellotti, and Gwizdka (2006) consolidate these into three main functions:

1. Task management;
2. Personal archiving; and
3. Contact management.

As noted earlier, our focus is on the first of these three functions: task management. We work to address limitations in existing email software by automatically detecting and identifying requests and commitments within email messages, to help users give appropriate attention to and track the real-world obligations that are expressed or implied by each email message that they send or receive.

Popular media have been quick to trumpet more recent communication trends and tools, including the use of social messaging platforms such as Facebook and Twitter. Claims are made that such tools threaten to break the communication dominance of email. Despite the rapid growth in use of alternate communication software, however, email remains the *de facto* standard for online communication. It is especially unclear whether these newer platforms will ever replace email for task management in the workplace. Even as communication continues to spill across different media and devices to other platforms and tools, predictions for business use of email suggest continued growth into the foreseeable future. Table 1.1 shows predictions of worldwide email volumes out to 2016, based on analysis by the Radicati Group (Hoang, 2012). Of particular interest for our work is the continued dramatic increase in email volumes for business use, with predicted cumulative growth of more than 60% between 2012 and 2016. This contrasts quite sharply with the predictions for personal email use, with predicted personal message volumes to fall by 13% over the same period. These numbers suggest that, at least in the near-term future, it is the personal rather than the business email usage that will be displaced by the alternate messaging platforms such as Twitter and Facebook.

[3] "Forget the World Wide Web, Internet telephony, streamed video, even online pornography. When it comes to the Internet's 'killer application', email is it." (Lai, 1996)

	2012	2013	2014	2015	2016	2012–2016
Total Emails Per Day (billion)	144.8	154.6	165.8	178.3	192.2	+47.4
Year-on-Year Change (%)		+7%	+7%	+8%	+8%	+33%
Business Emails Per Day (billion)	**89.0**	**101.0**	**114.3**	**128.6**	**143.8**	**+54.8**
Year-on-Year Change (%)		+13%	+13%	+13%	+12%	+62%
Personal Emails Per Day (billion)	55.8	53.6	51.5	49.7	48.4	−7.4
Year-on-Year Change (%)		−4%	−4%	−3%	−3%	−13%

TABLE 1.1: Predictions of worldwide daily email traffic for 2012–2016, according to The Radicati Group, Inc. (Hoang, 2012, p. 3).

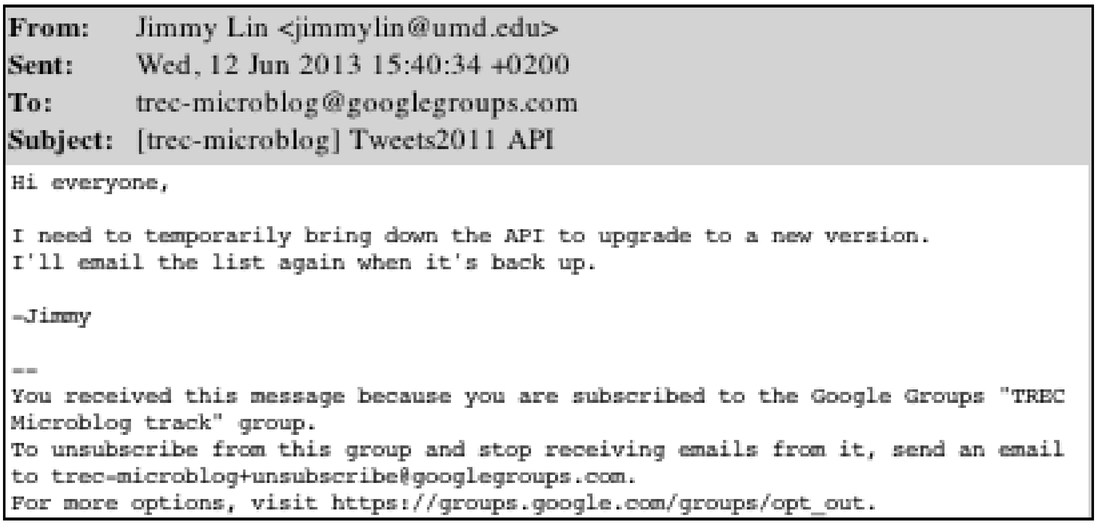
1.2 A Technical Overview of Email

At its heart, email is a computer-mediated communication system that allows people and systems to communicate by writing and exchanging textual messages. These messages are transmitted from the sender via a series of connected computers and computer networks to one or more designated recipients, where each receiver can open, read, interpret and potentially act on the message content. The format of, and method for transporting, each email message is defined in a series of Requests for Comment (RFCs) published by the Internet Engineering Task Force (IETF) (Resnick, 2008). These RFCs have evolved over time as technology and user requirements have changed, but remarkably, the core email standards remain almost unchanged from their earliest incarnation.

In terms of format, each email message is divided into lines of characters and consists of two main sections: the HEADER SECTION of the message, followed by the MESSAGE BODY. The header section contains a collection of metadata about the message; the message body contains the content of the message being sent. An example message showing these parts is given in Figure 1.1.

The header section is represented as a sequence of lines of text that contains structured and semi-structured HEADER FIELDS, and includes information about the sender, recipients, date, transport path from sender to recipient, metadata about the email software used to send the message, a globally unique message identifier, and often, details of spam detection, virus and malware scanning performed on the message by one or more email servers. This metadata is represented in the form of key–value pairs. Typically, an end-user email application will display only a handful of the available header fields, such as the sender, recipient list, date and subject line, as shown in Figure 1.1. Figure 1.2 shows the full header section for the email message in Figure 1.1.

We draw attention to this contrast between available and visible information to demonstrate that information that may be useful for computation, but not intended for a human reader, can be transmitted along with the message content without burdening the email recipient. This is important when we begin to think about ways to encode explicit information about the presence or location of request and commitment



From: Jimmy Lin <jimmylin@umd.edu>
Sent: Wed, 12 Jun 2013 15:40:34 +0200
To: trec-microblog@googlegroups.com
Subject: [trec-microblog] Tweets2011 API

```
Hi everyone,

I need to temporarily bring down the API to upgrade to a new version.
I'll email the list again when it's back up.

-Jimmy

--
You received this message because you are subscribed to the Google Groups "TREC
Microblog track" group.
To unsubscribe from this group and stop receiving emails from it, send an email
to trec-microblog+unsubscribe@googlegroups.com.
For more options, visit https://groups.google.com/groups/opt_out.
```

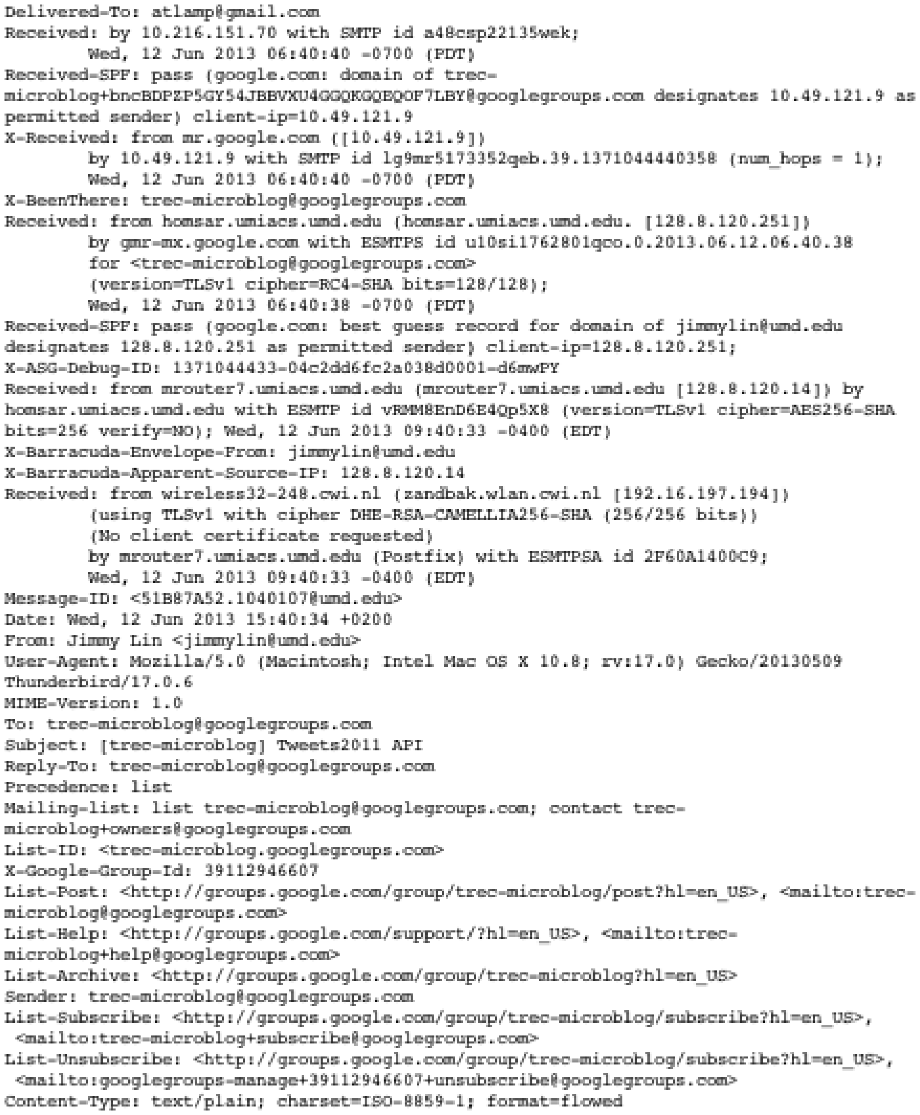
Figure 1.1: An email message showing the header section and body that make up each message.

utterances that occur within the body of a message, without interrupting the user's consumption of the email content itself. Such information can be unobtrusively stored within key–value pairs in the header section to be processed by task-aware email clients, and quietly ignored by legacy clients. We return to such issues when we discuss our prototype Microsoft Outlook integration software in Chapter 6.

The message body contains the free text content of the message. Originally, this was envisaged to include only textual content. Through the use of MIME (Multipurpose Internet Mail Extensions) standards (Freed and Borenstein, 1996), however, the body of an email message can also encapsulate any media types or binary data, including images, video, audio, documents and files, encoded as a stream of text. These capabilities have allowed email to become a common tool for the exchange of files and data, in addition to its intended communication role. That email has been successfully adopted for such originally unforeseen uses is a testimony to the flexibility and robustness of the original protocols and transport mechanisms.

1.3 The Problem of Email Overload

As we saw in Table 1.1, there were almost 145 billion emails sent every day in 2012. The sheer volume of incoming messages creates significant challenges for people using email. As early as the 1980s, researchers began to study how email was being employed in organisations. Denning (1982) was one of the first to write about the rise in electronic message volumes, and was an early advocate for the automated filtering of incoming messages. Sumner, another early researcher in this field, interviewed and surveyed



```
Delivered-To: atlamp@gmail.com
Received: by 10.216.151.70 with SMTP id a48csp22135wek;
        Wed, 12 Jun 2013 06:40:40 -0700 (PDT)
Received-SPF: pass (google.com: domain of trec-
microblog+bncBDPZP5GY54JBBVXU4GGQKGQEQOF7LBY@googlegroups.com designates 10.49.121.9 as
permitted sender) client-ip=10.49.121.9
X-Received: from mr.google.com ([10.49.121.9])
        by 10.49.121.9 with SMTP id lg9mr5173352qeb.39.1371044440358 (num_hops = 1);
        Wed, 12 Jun 2013 06:40:40 -0700 (PDT)
X-BeenThere: trec-microblog@googlegroups.com
Received: from homsar.umiacs.umd.edu (homsar.umiacs.umd.edu. [128.8.120.251])
        by gmr-mx.google.com with ESMTPS id u10si1762801qco.0.2013.06.12.06.40.38
        for <trec-microblog@googlegroups.com>
        (version=TLSv1 cipher=RC4-SHA bits=128/128);
        Wed, 12 Jun 2013 06:40:38 -0700 (PDT)
Received-SPF: pass (google.com: best guess record for domain of jimmylin@umd.edu
designates 128.8.120.251 as permitted sender) client-ip=128.8.120.251;
X-ASG-Debug-ID: 1371044433-04c2dd6fc2a038d0001-d6mwPY
Received: from mrouter7.umiacs.umd.edu (mrouter7.umiacs.umd.edu [128.8.120.14]) by
homsar.umiacs.umd.edu with ESMTP id vRMM8EnD6E4Qp5X8 (version=TLSv1 cipher=AES256-SHA
bits=256 verify=NO); Wed, 12 Jun 2013 09:40:33 -0400 (EDT)
X-Barracuda-Envelope-From: jimmylin@umd.edu
X-Barracuda-Apparent-Source-IP: 128.8.120.14
Received: from wireless32-248.cwi.nl (zandbak.wlan.cwi.nl [192.16.197.194])
        (using TLSv1 with cipher DHE-RSA-CAMELLIA256-SHA (256/256 bits))
        (No client certificate requested)
        by mrouter7.umiacs.umd.edu (Postfix) with ESMTPSA id 2F60A1400C9;
        Wed, 12 Jun 2013 09:40:33 -0400 (EDT)
Message-ID: <51B87A52.1040107@umd.edu>
Date: Wed, 12 Jun 2013 15:40:34 +0200
From: Jimmy Lin <jimmylin@umd.edu>
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10.8; rv:17.0) Gecko/20130509
Thunderbird/17.0.6
MIME-Version: 1.0
To: trec-microblog@googlegroups.com
Subject: [trec-microblog] Tweets2011 API
Reply-To: trec-microblog@googlegroups.com
Precedence: list
Mailing-list: list trec-microblog@googlegroups.com; contact trec-
microblog+owners@googlegroups.com
List-ID: <trec-microblog.googlegroups.com>
X-Google-Group-Id: 39112946607
List-Post: <http://groups.google.com/group/trec-microblog/post?hl=en_US>, <mailto:trec-
microblog@googlegroups.com>
List-Help: <http://groups.google.com/support/?hl=en_US>, <mailto:trec-
microblog+help@googlegroups.com>
List-Archive: <http://groups.google.com/group/trec-microblog?hl=en_US>
Sender: trec-microblog@googlegroups.com
List-Subscribe: <http://groups.google.com/group/trec-microblog/subscribe?hl=en_US>,
 <mailto:trec-microblog+subscribe@googlegroups.com>
List-Unsubscribe: <http://groups.google.com/group/trec-microblog/subscribe?hl=en_US>,
 <mailto:googlegroups-manage+39112946607+unsubscribe@googlegroups.com>
Content-Type: text/plain; charset=ISO-8859-1; format=flowed
```

FIGURE 1.2: The full set of header lines for the email message in Figure 1.1.

users at an organisation with an electronic email system in heavy use (Sumner, 1988); she found that email was displacing previous communication modalities and warned that access to electronic mail systems might lead to information overload, where users feel overwhelmed by the number of email messages they receive and frustrated by the amount of time they need to spend dealing with them.

To try to understand why email overload occurs, Thomas et al. (2006) reviewed the email logs of managers in an "acquisition program agency" within the US Department of Army. Based on an analysis of the data and interviews with staff, they identified key reasons for email overload, including that it was easy and cheap for users to broadcast their messages to many users, and that there was a perception of constant pressure to be available and responding to email, even while completing other tasks. Dabbish et al. (2005) have drawn a direct link between email overload and task management; they found that the impact of feelings of email overload includes a reduced ability to coordinate work. Fisher et al. (2006) also suggest that, as email continues to be used for all manner of information storage and communication, email overload will continue unabated.

Spam, or unsolicited commercial email (UCE), is another factor in email overload; Jones (2008) is one researcher who has highlighted how people feel particularly frustrated by the amount of this material they receive. In recent years, however, several studies have suggested that spam no longer causes the same level of pain for users as it once did. Research from the Pew Internet and American Life Project, for example, reveals that email users are increasingly using the filters offered by their email providers or employers to block spam (Fallows, 2007), and that, while the overall volume of spam being sent is increasing, surveyed email users are becoming less bothered by it. This is likely due to the more effective spam filtering techniques that have been developed and deployed on widely used email systems such as Gmail and Hotmail, as well as in corporate email systems. These are able to prevent most spam messages from reaching a user's inbox.

In summary, despite technological innovations reducing the burden of spam for many email users, the problem of email overload more generally remains a significant issue, and ongoing research suggests that feelings of email overload will continue to grow, particularly for those dealing with concurrent, task-related conversations in the workplace.

1.4 Using Email as a Task Management Tool

Email is widely used for task-oriented organisational communication, which has contributed to many users complaining of being overwhelmed by the volume of messages they receive. Despite its popularity as a means for exchanging, delegating and reporting on tasks, there remain significant problems in using email as a tool to manage and track tasks (Whittaker and Sidner, 1996; Bellotti et al., 2003). Despite research and wide personal experience that clearly demonstrate a range of difficulties that people face in executing and managing tasks via email, modern email clients have largely failed to adapt to or accommodate such uses. Indeed, it is widely acknowledged that current interfaces for managing email provide poor support for the task management activities for which users have adopted them (Hiltz and Turoff, 1985; Whittaker and Sidner, 1996; Ducheneaut and Bellotti, 2001). The result is that many users continue to have difficulty giving appropriate attention to the requests and commitments that lie hidden in their email requiring action or response. Studies of task-focused email usage repeatedly uncover problems with "keeping track of lots of concurrent actions: One's own to-dos and to-dos one expects from others" using existing email clients (Bellotti et al., 2003, p. 346).

Dabbish et al. (2005) determined that email users in their study wanted those messages that requested action to be kept visible in their email client, but struggled to achieve this with their available email software. The authors suggested that systems could be developed to make these important aspects of messages more visible, or to provide mechanisms to "disassociate the reminders from the message themselves".

Making important aspects of messages more visible can be addressed at the MESSAGE LEVEL. For our purposes, this requires tools and techniques to determine *whether* an email message contains requests or commitments. Extracting tasks or reminders from the messages themselves must be addressed at the UTTERANCE LEVEL; this requires an ability to determine precisely where and how each request or commitment is expressed. In this thesis, we address both these tasks, with the aim of creating tools and systems to assist email users by automatically detecting requests and commitments in incoming and outgoing email. This capability affords the types of interfaces requested by Dabbish et al.'s studied users, for example in the form of markers on messages containing requests or commitments, and through extracting request and commitment utterances to structured task lists.

1.5 The Contributions of this Thesis

The work described in this thesis makes several contributions to both understanding and automatically identifying requests and commitments in workplace email. The thesis:

- explores real-world usage of requests and commitments in workplace email and, based on analysis of actual data, presents comprehensive definitions that account for the complexity observed;
- develops and evaluates a series of classifiers for automatically identifying requests and commitments at three levels of granularity, with significantly better performance than baseline systems;
- proposes and evaluates a novel coarse-to-fine classifier ensemble to counter high levels of data skew, which reliably improves the performance of automated commitment identification at fine granularities;
- develops and evaluates a novel system for automatically identifying functional zones within email messages, with cross-validation accuracy of more than 90% in two-zone and three-zone configurations;
- introduces an annotated corpus of 1000 zone-annotated email messages that has been made publicly available for research use, under a Creative Commons Non-Commercial Attribution licence;[4]
- demonstrates that performing automated zoning prior to identifying requests or commitments in an email message improves classification accuracy by 5–16%;

[4] Available from: http://zebra.thoughtlets.org/data.php.

- creates and exploits a series of manually annotated email corpora, culminating in a span-annotated corpus of 1000 email messages that has been made publicly available for research use, under a Creative Commons licence;[5] and
- develops a prototype software plug-in for Microsoft Outlook that integrates message-level and sentence-level request and commitment classifiers into the world's most widely used business email software.

The combination of in-depth analysis, email annotation tools, automated classifiers and the software we have developed to integrate these into Microsoft Outlook provide a well grounded foundation and a tangible platform for other researchers to conduct further empirical task-related email research with real-world data and in real-world contexts.

1.6 The Structure of this Thesis

Our aim in this thesis is to create tools that assist email users to identify, manage and ultimately react or respond appropriately to requests and commitments in incoming and outgoing email. We focus on requests and commitments, as they have been shown to be the communicative obligation acts at the heart of task-based communication (Winograd and Flores, 1986). In order to build software tools that can recognise these acts, we first seek to understand how task-oriented communication is conducted via email in real-world interactions. This exploration begins in Chapter 2, where we provide an overview of related work, and review the linguistic theories underlying the work in this thesis. We focus on Speech Act Theory, a theory from the study of the philosophy of language that focuses on identifying and categorising the functional meaning of utterances. We examine how Speech Act Theory can be applied to email communication, and how others have categorised and attempted to automate the identification of different speech acts for utterances in spoken and written language.

Chapter 3 then outlines our analysis of how requests and commitments are made in email. As with any other pragmatic analysis of language, we face the challenge of needing to establish which language forms can be used to represent requests and commitments, a task we approach by gathering and analysing judgements from multiple human annotators. Our empirical observations are thus drawn from analysing a large collection of manually labelled requests and commitments in real-world email, assembled via a series of annotation experiments. From our analysis, it is clear that much previous work has overlooked the complexity involved in how people exchange requests and commitments in real-world email. This chapter provides the foundations for future task-oriented email research, by providing a systematic account of the complexities that must be considered when coding or classifying email text, along with a set of comprehensive definitions and annotation guidelines for recognising request and commitment acts.

[5] Available from http://zebra.thoughtlets.org/spans/data.php.

We then consider how to build effective tools and systems to automatically identify actionable content within email messages. Chapter 4 describes the system that we designed, built and evaluated for automating the detection of requests and commitments in email. This system operates at the message level, classifying each message according to whether or not it contains one or more requests or commitments. This chapter also describes the design, implementation and evaluation of a system called Zebra that we developed to automatically segment the body text of email messages into their different functional sections, a task we refer to as EMAIL ZONING. Our evaluation experiments provide another key contribution from this thesis: that knowledge of the locations of the different email zones within an email message improves the performance of request and commitment classifiers by allowing them to focus only on the relevant text within each message, ignoring spurious content such as advertising, email signatures, legal disclaimers and quoted content from previous messages.

Building on our message-level classification work, in Chapter 5 we then adapt our approach to automatically identify requests and commitments at finer levels of granularity, specifically at the paragraph and sentence level. This builds the foundations required to provide further assistance to people working with tasks in email, in the form of services such as extractive summarisation of actionable content or the automatic extraction of requests and commitments to external task management systems. This chapter demonstrates the additional difficulties that arise due to the highly skewed nature of data at this finer granularity. Despite these challenges, our work demonstrates that ensembles of fine- and coarse-grained classifiers working together outperform the same fine-grained classifiers working independently. This is an important result that has implications for other classification tasks with different hierarchical levels at which classification can be performed.

In Chapter 6, we discuss our experimental work to apply and integrate our automated message, paragraph and sentence classification systems into Microsoft Outlook, one of the most widely used software products for business email. The aim of this work is to increase the visibility and role of tasks within email clients, relative to the status they have in commonly used commercial email applications. We prototype how to expose message and sentence classifications within Microsoft Outlook and consider how these could be built upon to provide further user support such as action-oriented summaries of email messages and threads; task-based navigation and visualisations; and dashboards that provide overviews of the state of an email inbox or collection with much greater fidelity than is possible with current tools.

Finally, in Chapter 7, we reflect on the work presented in this thesis and point to opportunities for future experimentation and evaluation based on our research.

"Organizations exist as networks of directives and commissives"
Terry Winograd and Fernando Flores (1986, p. 157)

2 Literature Review

Despite its relatively short history, business, government and other organisations have grown to depend on email for workplace communication. As noted in Chapter 1, the volume of email messages sent and received has been increasing dramatically for many years, and email is used for an extremely wide variety of purposes from document distribution to task and contact management.

In this chapter, we dig more deeply into relevant work from the literature that has analysed workplace email usage and task-oriented communication. We begin in Section 2.1 by reviewing previous ethnographic, linguistic and computer science research that has examined workplace communication. Together, these studies have confirmed the importance of task-oriented communication in the workplace. More significantly, as we noted in Chapter 1, task-oriented email communication has been analysed and distilled down to the electronic exchange of directive and commissive communicative acts as a fundamental abstraction of the way work is delegated and completed within organisations. As the basis of task-oriented communication, we focus our research in this thesis on these OBLIGATION ACTS, which we term REQUESTS and COMMITMENTS.

In the context of this focus, we review Speech Act Theory, from which the notions of directive and commissive acts are drawn, in Section 2.2. Speech Act Theory, and other related work in linguistic pragmatics, seeks to analyse and identify the intentions and actions that lie behind utterances, and provides the domain-independent abstraction of task-oriented conversations that we seek to identify in email messages. In Section 2.3, we review previous work where Speech Act Theory has been applied to analysing email and other forms of computer-mediated communication (CMC). We look specifically at previous categorisations and definitions of speech and dialogue acts, with a view to re-using these for our own work. As our discussion makes clear, none of the surveyed categorisations meets our specific needs for robustly identifying request and commitment speech acts in email. Consequently, we adopt specific aspects and make use of insights, mistakes and successes from the design and application of the surveyed taxonomies to guide and inform our own analysis and to motivate the definitions of requests and commitments that we develop in Chapter 3.

Finally, in Section 2.4 we analyse and review experimental software systems that have been developed to assist users to work more efficiently with actionable content in email. Again, our goal in reviewing these systems is to identify promising approaches and insights, as well as gaps and opportunities, to improve support for people working with tasks in email. We explicitly build on our insights from these systems when we present our own prototype integration with Microsoft Outlook email software in Chapter 7.

This chapter lays the foundation for our analysis of the nature of requests and commitments in Chapter 3, as well as for our computational research that focuses on automatically identifying requests and commitments in email messages and how to present this information to users (Chapters 4, 5 and 7).

2.1 Characterising Workplace Email Communication

The nature of workplace communication, of which email is a highly significant part, has been well studied over many years. Much of this research has characterised face-to-face workplace communication as consisting of brief, frequent interactions and communication (Reder and Schwab, 1990; Kraut et al., 1992; Whittaker, Frohlich, and Daly-Jones, 1994). Rather than taking the time to batch up information to exchange with a colleague, we tend to repeatedly interact with them when we have new information to exchange with them or a new request to make of them. Whittaker et al. (Whittaker, Frohlich, and Daly-Jones, 1994; Whittaker and Schwarz, 1995) have analysed workplace conversations through a continuously recording video camera in an office environment and discovered similar patterns: communications were brief, unplanned, frequent and dyadic, that is, direct communication between two people or two groups of people.

Another consistent observation is that workers frequently engage in multiple separate interactions in order to complete a task. Complicating matters further, workers are also usually engaged in several concurrent conversations. They must track these separate conversational threads and switch context as appropriate (Whittaker and Schwarz, 1995). This suggests that tools designed for workplace communications should readily accommodate conversations that consist of many brief interactions between two or more participants.

Studies of email communication within organisations, for example (Whittaker and Sidner, 1996; Ducheneaut and Bellotti, 2001; Fisher et al., 2006), have uncovered communication patterns that reflect those observed for face-to-face communication, as outlined above. It is presumably for these reasons that most email software has, somewhat belatedly, added better support for allowing people to view and interact with email as conversations, rather than only as discrete messages.

Beyond these commonalities, however, email has also brought significant change, especially in terms of the volume and rhythms of organisational communication. Synchronous electronic media such as instant messaging, video conferencing and telephones allow participants to take part in the same conversation from different physical places, but they still require all people to be simultaneously available at some nominated time. As an asynchronous electronic medium, email frees participants from the constraints of *both* time and space, allowing senders and recipients to interact at times and in places that are convenient to each (Sproull and Kiesler, 1991). Together with technologies such as smart phones and Virtual Private Networks (VPNs) that have brought workplace email into people's personal time and space, the nature of email means that, in many organisations, the flow of workplace conversations continues around-the-clock, rather than abating when people physically leave the office.

Of course, there is a lot of organisational communication that does not neatly fit into the text-only, asynchronous, message-oriented communication for which email was originally designed. People also want to have synchronous or near-synchronous interaction that is richer and more conversation focused than the discrete interactions around which email was originally developed. They want to routinely exchange information (documents and other non-textual files) and manage their contact lists, not just take part in conversation. As Whittaker, Bellotti, and Moody (2005) note, the malleability of email has been a crucial factor in its popularity; email has adapted to many purposes which were not considered in its design, such as to operate as a task manager, to deliver and disseminate documents, and to create and maintain a contact management system (Mackay, 1988; Whittaker and Sidner, 1996; Bellotti et al., 2003). Increasingly, email is also used to provide a legally-recognised record of organisational interactions, as evidenced by the rapidly growing field of eDiscovery, which deals with the archiving, search and exploration of electronic organisational archives, primarily email repositories, for legal purposes (Baron, Lewis, and Oard, 2006).

As usage of email has broadened, email has naturally become the medium for a growing proportion of organisational communication. Ducheneaut and Bellotti have examined workplace email usage in some detail (2001; 2003), in particular looking at how and when users communicate via email in the workplace. Unsurprisingly, their research confirms that email is the major means of non face-to-face communication in the workplace, as well as a common means of document exchange. More surprising, however, is that even when collaborators work in plain sight of one another, they still send each other a substantial amount of email, suggesting that there are qualities of the email medium that make it the preferred medium for certain types of workplace communication, above face-to-face conversation. Ducheneaut and Bellotti also find that users are increasingly appropriating their email clients as a 'habitat' in which they spend most of their workday.

Despite its ubiquity in the workplace, however, email still has its problems. As we saw in Chapter 1, email overload remains a significant problem. Jackson, Dawson, and Wilson (2003) investigated the effect of email interruptions in the workplace, by electronically monitoring employees' activities. They found that the way the majority of users handle their incoming emails gives rise to far more interruptions than expected. This suggests that there is a place for email clients to do more to help users manage email interruptions in the face of ever-growing email volumes. Hair, Renaud, and Ramsay (2007) further observed that individual users respond differently towards email and potential email overload issues and developed a typology of user orientations or predispositions towards email based on the results of an online survey into email behaviours, as follows:

- Relaxed: email exerts no undue pressure. These users deal with emails as and when they see fit and experience email as an asynchronous communication medium.
- Driven: email exerts pressure. These users feel the need to reply almost instantaneously to emails and expect the same in return, and thus experience email as a near-synchronous communication medium.
- Stressed: email exerts stress. These users do not find email a useful medium, and in particular, find the pressure to respond to be a negative factor.

We believe that improving the efficiency and reducing the friction involved for users managing tasks in email has the potential to lower stress levels and increase productivity for all email users, regardless of how they respond to email and feelings of email overload.

2.1.1 The Challenges of Task Management in Email

As we touched on in Chapter 1, existing research has repeatedly confirmed that email is a popular tool for negotiating, delegating and reporting on tasks in the workplace (Mackay, 1988; Whittaker and Sidner, 1996; Ducheneaut and Bellotti, 2001; Whittaker, 2005; Fisher et al., 2006). A particularly lucid summary of how users would like to work with task-oriented content in email comes from work by Murray (1991), whose ethnographic research into the use of electronic messaging at IBM highlighted that:

> “[Managers] would like to be able to track outstanding promises they have made, promises made to them, requests they’ve made that have not been met and requests made of them that they have not fulfilled.”
>
> (Murray, 1991, p. 27)

Unfortunately, while this is how people would like to work with tasks in email, they frequently have difficulty organising and managing their email messages, and they experience significant problems in using email as a tool to manage collaborative tasks (Whittaker and Sidner, 1996; Bellotti et al., 2003). In the literature, it is widely acknowledged that current interfaces for managing email provide poor support for the task management activities for which users have adopted them (Hiltz and Turoff, 1985; Whittaker and Sidner, 1996; Ducheneaut and Bellotti, 2001).

For example, studies of email usage (Venolia et al., 2001; Whittaker, 2005) have shown that one very common and simple strategy for managing commitments in email is to respond or forward the original message to others who need to take action, and to leave the original message in the inbox as a reminder about that task. Users know that they will return to the inbox to access new messages, and hope that they will see the original message and recall the outstanding task. Although this strategy can work reasonably well as a reminder about requests that one receives, it is much less effective when trying to keep track of tasks that are delegated to others or reminding one about commitments made to others. Most email programs keep a copy of sent messages, but these are generally stored in a separate 'sent mail' folder, and users do not routinely see these when they access new email.

Having to remember to access sent mail makes it especially hard to keep track of actions that one is "owed", for example, from requests made in outgoing messages. One of Whittaker's study participants, a software support manager, describes exactly this challenge:

> "Well let's say that someone has asked me something. And I say I don't know. I will forward it to someone on my staff. Copy that person and say, "I need some help with this". And that's what, I will keep that sent message as a reminder. But the problem is that I still have to remember to open my sent mail to remind myself that they haven't replied yet."
>
> (Whittaker, 2005, p. 54)

The same issue arises for keeping track of actions that stem from commitments that one makes in email sent to others.

Some users attempt to resolve this particular issue by copying themselves on every sent message to generate inbox reminders, but this overloads the inbox still further, and simply shifts the problem back to dealing with an overflowing inbox. Gwizdka (2002) notes that email interfaces generally do not make it easy for users to track and manage future information and actions. This deficiency serves only to exacerbate the challenges faced by users who must repeatedly review messages left in their inbox for required action as time and memory permits.

There are users who manage to use their inbox effectively as a to-do list, capitalising on the fact that when they view incoming email, they are visually reminded of the majority of their outstanding tasks. This is, however, increasingly difficult given the ever-growing volume of email people receive. Many users seldom review their inbox for outstanding messages (Whittaker and Sidner, 1996). This task monitoring becomes all but impossible as users' inboxes become cluttered with email, because outstanding tasks are quickly displaced and scroll out of sight as new messages are received. Under such conditions, there are no longer visual cues to remind users of the state of their current conversations and tasks. With the volumes of workplace email experienced today, opportunistic reminding and task tracking within current email clients usually fails to keep up. This can easily lead to important tasks and actions being overlooked. In turn, an overloaded inbox leads to email users feeling overwhelmed and daunted by the time it takes to deal with all the incoming work (Hiltz and Turoff, 1985; Whittaker and Sidner, 1996).

Together, these challenges have motivated new research into tools and approaches that can assist people to more efficiently triage and more reliably manage tasks in their incoming and outgoing email messages. Later in this chapter, we look at a range of categorisation approaches that have been applied to identify and track tasks in email, as well as several experimental systems that have worked to address some of these issues. Our contribution to improving task management in email stems from the comprehensive approach we take to understanding the language of task-based email, which we use to inform the design, development and evaluation of new tools to automatically identify task-related content in email. In the next section, we look at the abstraction we use to analyse patterns of language use in task-based email, as the foundation of our approach.

2.2 Analysing Speech Acts in Email Communication

Email is an example of computer-mediated communication, and represents a "hybrid register that resembles both speech and writing and yet is neither" (Veselinova and Dry, 1995). Baron (1998) notes that many email users consider email to be "speech by other means", while still acknowledging their more formal messages to be more like "letters by phone". Email text combines written conventions with characteristics of spoken language in a form that continues to evolve into its own register (Danet, 1998). Since the register is relatively new in linguistic terms, "many email writers are drawing on their experience of both written and spoken communication when composing in this medium" (Daly, 1996). The dialogic nature of email also means that threads of email conversations can often share many similarities with conversational speech.

As we have already seen, email is used in wildly different ways for widely different purposes. Underlying many of these diverse uses, however, is a commonality of facilitating communication between two or more people. Given this, we follow in the footsteps of previous researchers and apply approaches and techniques used for linguistic analysis of spoken and written human discourse to email discourse.

2.2.1 Speech Act Theory

Our focus on identifying request and commitment utterances in email is based on concepts from Speech Act Theory, which originates from the study of the philosophy of language. Austin (1962) was the original author of Speech Act Theory, which was later refined by his student, Searle (1969; 1976; 1979; 1989). Speech Act Theory has been widely applied and adapted within various areas of linguistics and philosophy.

The primary aim of Speech Act Theory is to account for the functional meaning of utterances, beyond an interpretation of their propositional meaning. The core tenet of Speech Act Theory is that when someone speaks, they also act. Austin's work began from a dissatisfaction with the then commonly adopted, but clearly false, philosophical assumption that "the business of a [sentence] can only be to 'describe' some state of affairs, or to 'state some fact', which it must do either truly or falsely"(Austin, 1962, p.1). Austin observed that there are many utterances which on the surface appear to be fact-stating, but are actually quite different, and that cannot be assessed on their truth or falsehood. He focused on utterances such as the following:

(2.1) *You're fired!*

(2.2) *I quit!.*

In utterances such as these, the intention of the speaker is clearly not (primarily) to state a fact that may be true or false, but to effect some change on the world.

From this work emerged the term SPEECH ACT, which refers to what is done, as contrasted with what is said, when someone utters something. This is at the heart of Austin's insight that each utterance in a dialogue is a kind of action being performed by the speaker. For example, in the following utterance the speaker has just made a PROMISE:

(2.3) *I promise I will send you the document today*

A promise is one example of a speech act (though not all speech act schemes would identify it with exactly that label). Other common classes of speech acts are REQUESTS, ORDERS and QUESTIONS, which all attempt to place an obligation for action or response on the intended recipient.

Austin originally outlined his theory of speech acts by introducing the concept of PERFORMATIVE language. He created a clear distinction between performative utterances that change the state of the world, and what he called CONSTANTIVE utterances: statements that attempt to describe reality and can be judged true or false. Consider, for example, the utterances in Examples (2.1) and (2.2) above; these are both performative utterances. Another example is the utterance:

(2.4) *I promise that I will come.*

In this example, as in Examples (2.1) and (2.2), the speaker is performing the act of promising, as opposed to making a statement that may be judged true or false. In contrast, consider an utterance like the following:

(2.5) *That car is blue.*

This can be judged as being true or false in the context in which it was uttered. For most performative utterances, however, it does not make sense to ask whether the utterance is true or false, but rather whether it is FELICITOUS – that is, whether it is appropriate to the context in which it is uttered. Through the course of his work, Austin came to the conclusion that most utterances are in fact performative in nature, that is, the speaker is nearly always doing something by saying something. Today, speech acts are recognised as a widely applicable approach to modelling and investigating the pragmatics of utterances.

Austin and Searle distinguished three types of acts or effects that may be performed simultaneously when someone performs an utterance:

1. LOCUTIONARY ACTS: simply uttering words;
2. ILLOCUTIONARY ACTS: the action performed in making an utterance; and

3. PERLOCUTIONARY ACTS: some external effect on the actions or attitudes of others that occurs as a result of the utterance.

Sometimes a further distinction is made at the locutionary act level between semantically empty utterances such as *Oh!*, and propositional utterances, for example in the work of Jackson (1874) and Van Lancker (1987). The linguistic field of SEMANTICS focuses on the study of propositional utterances.

The illocutionary act performed in making an utterance captures the speaker's intention, and is the focus of the linguistic field of PRAGMATICS. The illocutionary act is referred to by Searle as the illocutionary force of a speech act. Depending upon the context, the words used, and the speaker's intention, a given utterance might be conventionally interpreted as a statement, a command, a question, a request, a promise, and so on. Sometimes, the illocutionary effect can be made explicit, as in Examples (2.1), (2.2) and (2.4). Such utterances are examples of the explicit performative utterances on which Austin originally focused. For other utterances, the illocutionary force of an utterance can be implicit or indirect, as we will see in Example (2.7). These form examples of INDIRECT SPEECH ACTS, which we discuss further in Chapter 3.

The perlocutionary act, or perlocutionary effect, associated with an utterance is the effect it has on the thoughts, beliefs or actions of the hearer. Consider the following two utterances:

(2.6) *I'm asking you to close the window.*

(2.7) *It's cold in here.*

The illocutionary force behind both these utterances might be a request to close the window. The intended perlocutionary effect might be to actually have the hearer walk across the room and close the window. As discussed above, the illocutionary effect can be implicit, and interpreted by convention and context, as in Example (2.7), or explicit, as in Example (2.6). The perlocutionary effect of an utterance cannot, however, be made explicit; it is not possible to felicitously utter an utterance like *I convince to you close the window* in the course of normal conversation.

Speech Act Theory, as developed by Searle, focuses on categorising and identifying the illocutionary force behind each utterance. For the work in this thesis, we take a similar focus, seeking to identify utterances where an email sender intends the illocutionary effect to be either directive or commissive. As we have explained earlier, we adopt the terms request and commitment respectively to refer to these classes of speech acts.

2.2.2 Dialogue Acts

While Austin and Searle's speech acts are a useful and widely used characterisation of one kind of pragmatic force, other approaches have built on the ideas in order to model a greater number of conversational functions that an utterance can play. These enriched speech acts are often called DIALOGUE ACTS.

While the term is frequently used with different degrees of formality, a dialogue act generally has two main components: a semantic content and a communicative function (Bunt et al., 2010). The semantic content specifies the objects, relations, actions, events that the dialogue act is about; the communicative function refers to the speech act performed.

Bunt (1994) describes dialogue acts as functions that update the dialogue context. He distinguishes between several types of context:

- the linguistic context, which refers to the surrounding utterances;
- the semantic context, which refers to the underlying task and domain;
- the physical context, which refers to the location of the participants and the interaction in time and space;
- the social context, which refers to the type of situation and roles the participants play with respect to each other, including rights and obligations; and
- the cognitive context, which refers to the mental states of the participants.

In addition, Bunt differentiates between task-oriented acts and dialogue control acts. All acts affect the linguistic and cognitive context, but the former also changes the semantic context, while dialogue control acts affect the social or physical context. More concretely, task-oriented acts concern the actual task being discussed, and include acts such as requests, questions, answers and instructions; dialogue control acts are focused on the communicative task surrounding the actual task under discussion, and include acts such as acknowledgements and turn-taking signals that deal with the meta-task of maintaining the dialogue between the participants.

Ultimately, the main distinction between speech acts and dialogue acts is that speech acts usually represent one dimension of a more wide-ranging dialogue act classification. The other dimensions in dialogue act modelling attempt to capture information about the interaction in which the speech act occurs. Specifically, dialogue act modelling attempts to capture aspects such as whether the act is responding to a previous act or looking for a future act in response, as well as information about the physical and social setting in which the interaction occurs.

We restrict our work to identifying the illocutionary force of utterances in email. In future work, it could be interesting to extend our work to examine the forward and backward looking connections between utterances, to piece together the connections between task-oriented utterances over the course of email conversations. Such work would build upon the early work done by Winograd and Flores, looking at conversations for action (Winograd and Flores, 1986; Winograd, 1987). This is also the path followed by more recent research, such as that of the Semanta system (Scerri et al., 2009; Scerri et al., 2010).

2.3 A Review of Speech Act Taxonomies

As noted briefly in Chapter 1, the electronic exchange of request and commitment speech acts has been identified as a fundamental basis of the way work is delegated and completed within organisations; from their work on the Coordinator system, Winograd and Flores concluded that

> "Organisations exist as networks of directives and commissives"
> (Winograd and Flores, 1986, p. 157)

By this, Flores and Winograd mean that the exchange of requests and commitments is how tasks actually get distributed and acted upon within organisations. From an organisational perspective, staff are involved in a network of conversations, both internally and with external parties such as customers, suppliers and partners. The organisation actually achieves outcomes such as selling goods or providing services through requests and commitments that are made and received in conversations to fulfil obligations.

Given the central importance of requests and commitments, our goal is to find reliable ways to identify these speech acts within task-based email conversations, and to draw these to people's attention as a way of making tasks a more central aspect of email software. In this section, we thus review previous definitions and categorisations of speech acts from the literature, particularly where these have been applied to or are derived from analyses of computer-mediated communication (CMC) such as email.

While investigating particular surface forms of language is relatively unproblematic, it is widely recognised that it is very difficult to establish a definitive set of surface forms that correlate with a particular speech act (Archer, Culpeper, and Davies, 2008, p. 614). Winograd and Flores (1986, p. 159) go further, and posit that it is actually impossible to formulate a precise correspondence between the surface forms of utterances in a conversation and the structure of obligations that are conveyed. This is largely because, in the normal course of human communication, the complete intent of any piece of discourse cannot be determined by lexical or semantic analysis alone, but must be determined by context. This is especially true for INDIRECT SPEECH ACTS, which are particularly common in email text (Hassell and Christensen, 1996). We explore indirect speech acts further in Chapter 3, but briefly, they are utterances where a literal interpretation of the propositional content does not capture the speaker's desired illocutionary effect. Consider the following utterance:

(2.8) *I'd love a drink.*

Conventionally, this would be intended as a request for a drink, but this interpretation requires a non-literal interpretation. The literal interpretation would be that the speaker is simply revealing that he likes drinking or that he would be happy if he happened to have a drink. For this reason, under the request interpretation, this utterance is an example of an indirect speech act.

The point to be made here is that the complexity and context-sensitivity associated with identifying speech acts, especially indirect speech acts, leads to subjectivity, and, as we will see below, this in turn leads to significant variation in the taxonomies, categories and definitions that we survey. Unfortunately, despite the wide variety of taxonomies that we survey below, none are a good match for our needs in reliably identifying requests and commitments in real-world email. Instead, as we elaborate on in Chapter 3, we derive our own definitions based on lessons from this collection of taxonomies and definitions, coupled with insights from our empirical annotation of real-world email messages.

2.3.1 The Verbal Response Modes Taxonomy

Verbal Response Modes (VRM) (Stiles, 1992) is a principled taxonomy of speech acts that can be used to classify the illocutionary force of utterances. Each utterance is coded twice, once for its literal meaning, and once for its communicative intent or pragmatic meaning. The same eight VRM categories are used in each case.

The VRM categories are based on a principled division of utterances along three binary dimensions:

1. Source of Experience: The experience that a speech act concerns may be the speaker's own, as when the speaker reveals his/her own feeling or opinions, or it may be some other person's experience, as when the speaker asks a question or describes some other's feelings.

2. Presumption About Experience: In performing any speech act, the speaker may or may not need to presume to know what the other's experience was, will be or should be.

3. Frame of Reference: The meaning that an experience has in a particular speech act derives from the associated ideas, memories, connotations etc. with which it is linked. This is called the frame of reference of the speech act, and may be coded as either the speaker's or some other's viewpoint.

To elaborate on these dimensions, under the VRM system, every utterance from one person to another can be considered to concern either the *speaker's* or some *other's* experience. For example, consider the utterance:

(2.9) *I play with Lego.*

In this case, the speaker is talking about their own internal belief, so the source of experience is the *speaker*. In contrast, consider the utterance:

(2.10) *Do you play with Lego?*

In this case, the speaker is posing a question that seeks to understand the hearer's internal belief, so the source of experience is marked as *other* (non-speaker).

Further, in making an utterance, the speaker may or may not need to make presumptions about the other's experience. For example, in the utterance shown above in

Example (2.10), the speaker does not need to presume to know what the other person is, was, will be, or should be thinking, feeling, perceiving or intending. Such utterances require only a presumption of experience of the *speaker*. In contrast, consider the utterance:

(2.11) *Play with Lego!*

As a directive act, this utterance attempts to impose an experience (playing with Lego) on the hearer, and thus is marked as *other* for the presumption of experience.

Finally, in producing an utterance, a speaker may represent the experience either from their own personal point of view, or from a viewpoint that is shared or held in common with the other. The utterances shown above in Examples (2.9), (2.10) and (2.11) all use the speaker's frame of reference, because the experience is understood from the speaker's point of view. In contrast, the utterance:

(2.12) *You play with Lego.*

takes the other's frame of reference, representing the experience as the other views it.

These three principles—source of experience, presumption about experience and frame of reference—form the basis of the VRM taxonomy. The principles are dichotomous—each can take the value of *speaker* or *other*—and thus define eight mutually exclusive VRM categories, as shown in Table 2.1.

VRM applies two categories for each utterance classification: one to classify grammatical form and the other pragmatic meaning. This neatly handles the case of indirect speech acts which are not modelled as explicitly in many of the other taxonomies surveyed: utterances where the grammatical form and pragmatic meaning share the same VRM category are instances of direct speech acts; those where these categories differ are indirect speech acts. Stiles's VRM taxonomy therefore theoretically allows for sixty-four types of speech acts (each of the eight VRM categories in combination with every other VRM category). In practice, however, many of these combinations of grammatical and pragmatic VRM classifications are either uncommon or implausible in real-world discourse, so the number of observed VRM categories for real-world data is considerably less than this.

To validate the VRM taxonomy of speech acts, several hundred annotators have used it to classify several million utterances (Stiles and White, 1981; Miller and Stiles, 1986; Rak and McMullen, 1987; Hinkle, Stiles, and Taylor, 1988; Kline, Hennen, and Farrell, 1990; Meeuswesen, Schaap, and van der Staak, 1991; Stiles, 1992; Lampert, Dale, and Paris, 2006; Goldspink, 2007), making it almost certainly the most widely used speech act taxonomy that we survey. Surprisingly, however, we found no evidence that the VRM taxonomy had been used within the computational linguistics or natural language processing communities. We can only surmise that it has been overlooked because it was developed and has been applied in fields that fall outside the usual spheres of influence for NLP and CL work.

We began our work in this thesis using VRM as our initial taxonomy. As best we can tell, we were the first to automate the coding of VRM categories for utterances

Source of Experience	*Presumption about Experience*	*Frame of Reference*	*VRM Mode*	*Description*
Speaker	Speaker	Speaker	Disclosure (D)	Reveals thoughts, feelings, perceptions or intentions.
		Other	Edification (E)	States objective information.
	Other	Speaker	Advisement (A)	Attempts to guide behaviour; suggestions, commands, permission, prohibition.
		Other	Confirmation (C)	Compares speaker's experience with other's; agreement, disagreement, shared experience or belief.
Other	Speaker	Speaker	Question (Q)	Requests information or guidance
		Other	Acknowledgement (K)	Conveys receipt of or receptiveness to other's communication; simple acceptance, salutations.
	Other	Speaker	Interpretation (I)	Explains or labels the other; judgements or evaluations of the other's experience or behaviour.
		Other	Reflection (R)	Puts other's experience into words; repetitions, restatements, clarifications.

Table 2.1: The taxonomy of Verbal Response Modes.

(Lampert, Dale, and Paris, 2006), work that has since been built upon by a range of other researchers in the Natural Language Processing community—e.g., (Mildinhall and Noyes, 2008; De Felice and Deane, 2012). After that initial work, however, we found challenges in aligning our interest in task-oriented request and commitment speech acts with the VRM categories. Requests were reasonably consistent, frequently occurring as Advisements or Questions in terms of their pragmatic meaning. Commitments, however, were less consistent. Some commitments, such as:

(2.13) *Bob will send the document later today*

would be classified as Edifications, but this fails to distinguish them from the many other utterances that have no commissive intent, such as the utterance in Example (2.12), that would also be labelled as an Edification. Other commitments like:

(2.14) *I'll do it*

would be marked as Disclosures, but again, they share this category with many other non-commissive acts such as:

(2.15) *I like green eggs and ham*

As a result of this inability to reliably identify and distinguish commitments, we moved away from using the VRM taxonomy, and thus do not make use of it in the rest of the work in this thesis.

2.3.2 The VerbMobil Taxonomy

The first speech act taxonomy that we review is the VerbMobil scheme (Alexandersson et al., 1998), which was developed for use in annotating speech acts in two-party scheduling dialogues within the VerbMobil project (Wahlster, 2000). The speech act scheme used in the VerbMobil project contained a hierarchy of thirty-two acts which are sometimes grouped into three sets: one set describes control of the dialogue (e.g., GREET and BYE which are used to handle social obligations, rather than deal with a specific task); the second deals with management of the task conversation (e.g., INIT for initiating a dialogue segment that deals with a task and DEFER for deferring the dialogue segment that deals with a task); and the third manages solving or getting on with the actual task (e.g., REQUEST and COMMIT actions). This hierarchy of acts, grouped into these three categories, is shown in Figure 2.1.

The VerbMobil hierarchy works as both a convenient structure for understanding the available speech acts and as an annotation decision tree. When coding an utterance, the annotator navigates the tree in a top-down fashion, answering a question at each branching node to decide which branch to follow. If an answer cannot be given, traversal stops, and the label at the current node is used for annotation. Thus, utterances can be labelled with both leaf and non-leaf nodes from this tree. Every bold label in the tree shown in Figure 2.1 is a legal annotation tag, so it is valid to label a speech act with a REQUEST tag, rather than the more specific REQUEST SUGGEST, but it is not valid to mark any speech act with a PROMOTE TASK, MANAGE TASK, CONTROL DIALOGUE or DIALOGUE ACT label.

The work on speech acts for VerbMobil was one of the largest early speech act annotation experiments undertaken, with 76210 dialogue acts coded across German, English and Japanese dialogues.[1] Its primary focus was on exploiting knowledge of dialogue acts to improve machine translation of spoken dialogues. Based on this work, the VerbMobil taxonomy was an influential starting point for later work on the DAMSL standardisation effort (Core and Allen, 1997), which we will discuss in Section 2.3.3.

[1] These statistics are published at http://verbmobil.dfki.de/facts.html.

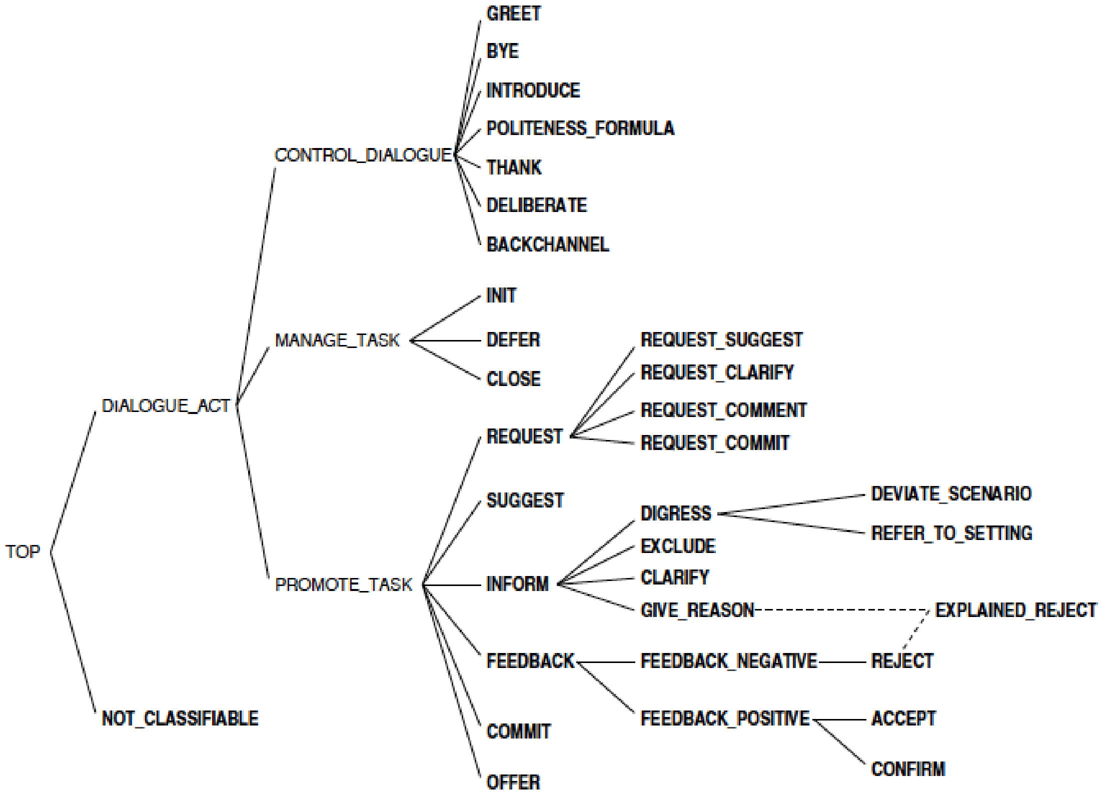


Figure 2.1: The VerbMobil taxonomy of speech acts.

The VerbMobil taxonomy has excellent processes defined for controlling and guiding the annotation process, and for training human annotators. The taxonomy has also been developed over a multi-lingual corpus of conversations and used to annotate dialogues in English, German and Japanese. It is therefore demonstrably language independent.

Despite these appealing characteristics, however, the taxonomy has, to our knowledge, only been applied to data in the travel planning and appointment scheduling domains which were the focus of the VerbMobil projects. As such, the speech act categories reflect the language used in scheduling, and have not been shown to be applicable to other domains of discourse. The VerbMobil researchers themselves acknowledge that their taxonomy is very useful in their domains of focus, but does not cover all aspects of natural dialogues that might occur across other domains (Alexandersson et al., 2000, p. 451).

There can also be ambiguity in some of the distinctions drawn by the VerbMobil categories. Consider, for example, the following brief exchange:

(2.16) Alice: *Shall we meet on the 27th?*

Bob: *The 27th sounds fine.*

Bob's utterance can be legitimately coded as either a Commit or a Feedback Positive, depending on whether we consider that Bob is simply acknowledging and accepting a suggestion, or committing to future action. In some contexts, VerbMobil actually prefers to categorise Bob's utterance with Accept Date, a more specific form of Feedback Positive label. The Accept Date category feels more like a semantic category than a pragmatic one, that is, it seems to capture the content of what is being discussed, rather than the communicative function. Our aim is to capture only pragmatic distinctions in our speech act taxonomy, to allow task-related speech acts to be recognised across many domains of email conversation.

More pragmatically, despite making approaches to several researchers who had been involved in the project, we were unable to identify any publicly available, annotated data that we could use to better understand the annotation outcomes or use as training or testing data for human annotators or algorithms. We have certainly drawn insights from the detailed and well-grounded VerbMobil taxonomy and annotation guidelines; however, based on the points discussed above, we do not directly adopt them in our own work.

2.3.3 The DAMSL Taxonomy

The Dialogue Act Markup with Several Layers (DAMSL) taxonomy (Core and Allen, 1997) is one of the most well-known dialogue act coding schemes. It was developed by the Discourse Resource Initiative (DRI) as a joint effort between several dialogue research groups, including those involved in the VerbMobil project, to standardise a single coding scheme for dialogue acts. Initially, DAMSL was focused on specific domains of task-oriented dialogues, though its focus later extended to developing a generic, domain-independent dialogue annotation scheme.

The DAMSL annotation scheme is divided into four layers of description for each utterance:

1. Communicative Status: Records whether an utterance is intelligible and whether it was successfully completed;

2. Information Level: A characterisation of the semantic content of an utterance;

3. The Forward Looking Function: How the current utterance constrains the future beliefs and actions of the participants, and how it affects the discourse; and

4. The Backward Looking Function: How the current utterance relates to the previous discourse.

In this structure, we see the wider-scoped characterisation of dialogue acts that we discussed back in Section 2.2.2. Only two of these levels are really relevant for speech act annotation: the forward and backward looking functions. The forward looking function corresponds to the illocutionary force of an utterance. The backward looking function draws on concepts from the field of conversation analysis, and specifically on the notion of adjacency pairs, focusing on the relationship of an utterance to previous utterances by the other speaker. Adjacency pairs are sequential pairs of utterances that commonly occur in dialogue, such as Question and Answer or Propose and Accept.

Adjacency pairs capture the tendency for an utterance of the first type (e.g., Question) to create an expectation for an utterance of the second type (e.g., Answer) to be issued in response. While a hearer is not strictly required to produce the expected speech act in response, they must be ready to justify their response if they choose to diverge from the expected response, and to accept responsibility for any inferences that the speaker might make as a result (Schlegloff and Sacks, 1973). The backward looking function of a DAMSL act attempts to capture these adjacency relationships between speech acts.

It is worth noting, however, that adjacency pairs capture only one form of sequencing in dialogue, and there are many interactions that do not conform to this pairing structure. Schlegloff (1972) provides examples of other specific dialogue sequences that are not adjacency pairs, such as rhetorical questions. Tsui (1989) argues that a three-part exchange is a more adequate basic unit of conversation analysis than adjacency pairs, due to the large number of conversation units that are observed that use the three-part structure; consider, for example, the following:

(2.17) Alice: *Could I have that please?*
Bob: *Sure*
Alice: *Thanks*

While we have only touched on the topic here, determining the relationship between utterances is clearly a complex undertaking; for our purposes of identifying task-related speech acts in email, we restrict our scope to identifying the communicative function of utterances. In DAMSL terms, this means that our focus is limited to the forward looking functions of utterances.

The forward looking communicative functions proposed by DAMSL are as follows:

- Statement
  - Assert
  - Reassert
  - Other-statement
- Influencing-addressee-future-action
  - Open-option
  - Action-directive
- Info-request
- Committing-speaker-future-action

  - Offer
  - Commit
- Conventional Opening Closing
- Explicit-performative
- Exclamation
- Other-forward-function

The most relevant categories for requests are the categories under INFLUENCING-ADDRESSEE-FUTURE-ACTION and INFO-REQUEST; for commitments, the relevant category is the COMMITTING-SPEAKER-FUTURE-ACTION category, which includes the OFFER and COMMIT sub-categories.

One issue with DAMSL's communicative functions is that the decision trees that guide annotators sometimes require the annotator to see how the listener responds to the speaker to determine what the speaker's actions were, and then annotate a particular utterance. This approach works well for retrospective analysis of conversations, but cannot be applied when annotating in-progress and incomplete dialogues, as we aim to do when detecting task-related utterances in email messages. This is clearly more of an issue for asynchronous email conversations than synchronous spoken conversations, where the gap between speaker turns is ordinarily very brief.

Finally, due to the generality and comprehensiveness of DAMSL, several researchers have argued that DAMSL is impractical for use in training a statistical classifier. To assess the complexity of DAMSL coding, consider the dialogue fragment, drawn from the TRAINS corpus, shown in Example (2.18):

(2.18) Alice: *Can I help you?*
Bob: *yes*

In Example (2.18), Bob's single-word utterance is annotated as:

- Agreement=Accept
- Answer=Yes
- Influence-on-speaker=Commit
- Info-level=Communication-management
- Response-to="utt2"
- Speech="-s 1.71275 -e 2.25275"
- Statement=Assert

As is clear from this annotation, there is significant complexity in annotating even short, simple utterances. Stolcke et al. (2000) attempted to use DAMSL to label dialogue acts in human-to-human telephone dialogue and found the taxonomy too unwieldy for practical use. In particular, they found DAMSL too complex for their aim of building a statistical classifier that could automatically categorise utterances according to their embodied dialogue acts.

Although there are only four levels of utterance annotations, some levels (particularly the forward and backward looking functions) have multiple dimensions that can be coded. With eight dimensions in the forward looking function, and four in the backward looking function, in addition to the four levels of annotations, there are around four million possible combinations of DAMSL tags (Clark and Popescu-Belis, 2004), which make for a massive space of possible annotations. This is problematic for both human annotator agreement and automated classification.

Despite these issues, the DAMSL speech acts are generally carefully defined, and we borrow from these definitions for scoping our own definitions and examining real-world requests and commitments in Chapter 3.

2.3.4 Camino's Taxonomy

Camino et al. (1998) analysed requests and corresponding answers in the context of exploring the efficacy of structured email messages. They defined a message-level annotation scheme which was used to classify requests based on the form of the expected answer. Their set of expected answer forms is:

- One-of-list choice, where the sender provided the recipient with two or more choices, from which the recipient was expected to choose only one;
- Several-of-list choice, where the sender provided the recipient with two or more choices, from which the recipient was expected to choose one or more;
- Free Text, where the answer to a request was judged to require explanatory text, or where choices were unlimited;
- Date, where the sender did not provide any explicit choices, and requested a date in response;
- Time, where the sender did not provide any explicit choices, and requested either an hour of the day or an explicit time;
- Number, where the sender requested a number not satisfying one of the above categories; and
- Exception request, where the sender requested that the recipient acknowledge some state (e.g., that the sent information was useful, received, or correct).

Camino et al. also distinguished requests requiring an email response—"a question or command or some other directive for which a response was important or could reasonably be expected" (Camino et al., 1998, p. 765)—from those requiring a physical action. Under these definitions, the request:

(2.19) *When is the conference deadline?*

would be considered a "request for an email act", requiring an answer via email, while a request such as:

(2.20) *Let's talk about this in my office, now* would be classed as a "request for a non-email act".

Camino's taxonomy is generally well considered for requests, but does not cater for real-world usage of conditional requests. The authors note that "Exception requests were considered unique because they are conditional: recipients are not expected to reply at all unless some exception state existed" (Camino et al., 1998, p. 766). This definition is the only one that covers conditionality in requests, and thus restricts conditional requests to being EXCEPTION REQUESTS. This, however, makes it unclear how to label common utterances such as the following:

(2.21) *If you'd like to attend, please let me know which date suits you best*

While the utterance in Example (2.21) clearly conveys a conditional request, the request does not fit the EXCEPTION REQUEST definition, nor any of the other defined request types. As we demonstrate in Chapter 3, such conditional requests are very common in email communication.

Camino's taxonomy also only caters for requests; there are no suitable categories for identifying commitments. More generally, the taxonomy is clearly also very focused on form-based aspects of classification, rather than distinguishing requests based on their illocutionary force. This focus does not align well with our needs for identifying requests and commitments in open domain email text, given the loose coupling between form and function that we have discussed earlier in this chapter.

2.3.5 Khosravi and Wilks' Taxonomy

Khosravi and Wilks (1999) were among the first researchers to attempt to automate the detection of speech acts in email. We consider their work to be both pragmatic and innovative. Later in this chapter, in Section 2.4.2.1, we discuss Pyam, their prototype software implementation that was built to embody a rule-based approach to identifying the acts in their taxonomy.

Their work focuses on recognising requests in email messages from a university help desk. They identify three classes of requests:

1. REQUEST-ACTION in which the sender tries to get the receiver to take action (e.g., *Please close the door*);

2. REQUEST-INFORMATION in which the sender is seeking information (e.g., *What is the time?*); and

3. REQUEST-PERMISSION in which the sender asks for permissions to have some information or to perform some action (e.g., *Can I put files on the group disk area?*).

From these three classes, the authors posited ten speech act categories that combined these acts, presumably based on utterances that were empirically observed in their email corpus:

1. REQUEST-INFORMATION;

2. Yes/No question and Request-Information;
3. Wh-Question and Request-Information;
4. Literally Yes/No Question and Indirect Request-Information;
5. Request-Action;
6. Literally Yes/No Question and Indirect Request-Action;
7. Request-Permission;
8. Literally Yes/No Question and Indirect Request-Permission;
9. Request Permission or Request Action; and
10. Request Permission or Request Information.

As we saw with Camino's taxonomy, many of the categories and distinctions drawn are syntactic rather than functional (e.g., the difference between a *yes/no* question and a *wh*-question). This syntactic focus of the categorisation also applies at the high level, in that each class of request (action/information/permission) is defined in terms of predetermined surface forms, rather than by the sender's intention. The authors acknowledge this, noting that their approach is "not a strongly theoretical approach for identifying speech acts" (Khosravi and Wilks, 1999, p. 249). As with Camino's taxonomy, Khosravi and Wilks' taxonomy also does not accommodate commissive speech acts, which limits its direct application for use in our work.

2.3.6 The DAMSL Switchboard Taxonomy

To address the limitations imposed by the complexity of working with DAMSL, Stolcke et al. (2000) derived their own smaller taxonomy from DAMSL for use in their Switchboard project. Before deciding to create their own taxonomy, a collection of utterances from the Switchboard corpus were first annotated with DAMSL tags. They observed that only about 220 combinations of tags occurred in their data. These 220 labels were clustered into 42 tags to further reduce the search space. This derived taxonomy is referred to as SWBD-DAMSL or the DAMSL Switchboard taxonomy, after their corpus of telephone switchboard conversations.

The set of acts in the DAMSL Switchboard taxonomy is shown in Table 2.2. The three most frequent acts from the taxonomy that occur in the Switchboard data are statement, backchannel/acknowledge and opinion, which together make up 68% of all acts. Note that none of these are task-related acts. The most frequent acts that are relevant to actionable content are the yes-no questions, which represent less than 2% of acts in the Switchboard corpus.

The DAMSL Switchboard taxonomy focuses on tags that are both 'linguistically interesting' and readily identifiable from surface forms (with the intention of making the problem of automatically identifying such acts more feasible). The resulting DAMSL

Tag	Examples	%
STATEMENT	Me, I'm in the legal department	36%
BACKCHANNEL/ACKNOWLEDGE	Uh-huh	19%
OPINION	I think it's great	13%
ABANDONED/UNINTERPRETABLE	So, -/	6%
AGREEMENT/ACCEPT	That's exactly it	5%
APPRECIATION	I can imagine	2%
YES-NO-QUESTION	Do you have to have any special training?	2%
NON-VERBAL	⟨Laughter⟩, ⟨Throat-Clearing⟩	2%
YES ANSWERS	Yes	1%
CONVENTIONAL-CLOSING	Well, it's been nice talking to you	1%
WH-QUESTION	What did you wear to work today?	1%
NO ANSWERS	No.	1%
RESPONSE ACKNOWLEDGEMENT	Oh, okay	1%
HEDGE	I don't know if I'm making any sense or not	1%
DECLARATIVE YES-NO-QUESTION	So you can afford to get a house?	1%
OTHER	Well give me a break, you know	1%
BACKCHANNEL-QUESTION	Is that right?	1%
QUOTATION	You can't be pregnant and have cats	0.5%
SUMMARIZE/REFORMULATE	Oh, you mean you switched schools for the kids	0.5%
AFFIRMATIVE NON-YES ANSWERS	It is.	0.4%
ACTION-DIRECTIVE	Why don't you go first?	0.4%
COLLABORATIVE COMPLETION	Who aren't contributing?	0.4%
REPEAT-PHRASE	Oh-fajitas	0.3%
OPEN-QUESTION	How about you?	0.3%
HOLD BEFORE ANSWER/AGREEMENT	I'm drawing a blank	0.3%
RHETORICAL-QUESTIONS	Who would steal a newspaper?	0.2%
NEGATIVE NON-NO ANSWERS	Uh, not a whole lot	0.1%
SIGNAL-NON-UNDERSTANDING	Excuse me?	0.1%
OTHER ANSWERS	I don't know	0.1%
CONVENTIONAL-OPENING	How are you?	0.1%
OR-CLAUSE	or is it more of a company?	0.1%
DISPREFERRED-ANSWERS	Well, not so much that	0.1%
3RD-PARTY-TALK	My goodness, Dianne, get down from there	0.1%
OFFERS, OPTIONS & COMMITS	I'll have to check that out	0.1%
SELF-TALK	What's the word I'm looking for	0.1%
DOWNPLAYER	That's all right	0.1%
MAYBE/ACCEPT-PART	Something like that	< 0.1%
TAG-QUESTION	Right?	< 0.1%
DECLARATIVE WH-QUESTION	You are what kind of buff?	< 0.1%
APOLOGY	I'm sorry.	< 0.1%
THANKING	Hey thanks a lot	< 0.1%

TABLE 2.2: The taxonomy of 42 DAMSL Switchboard speech acts.

Switchboard taxonomy has the advantage of a significantly smaller set of tags. This facilitates better inter-annotator agreement and larger collections of data for each tag type.

In focusing on distinctions that are evident in surface forms, however, the DAMSL Switchboard taxonomy draws a range of category distinctions that are not relevant for our focus on domain-agnostic identification of task-related speech acts. While both the WH-QUESTION and a YES-NO QUESTION, for example, are relevant as task-related speech acts, the distinction between these categories is one grounded in form rather than the functional or pragmatic differences that our work seeks to identify. There are also a range of acts in the DAMSL Switchboard taxonomy that, while interesting to analyse in email communication, fall outside our scope of identifying task-related content. In particular, given the project's focus on conversational speech, there is a certain bias towards informal, conversational categories which reflect that genre. The categories in DAMSL Switchboard are also defined primarily by canonical examples, rather than by means of a comprehensive definition of the scope and edge cases of each category. In combination, these factors led us to not adopt the DAMSL Switchboard taxonomy for our analysis and detection of request and commitment acts in email.

2.3.7 The Meeting Recorder Dialogue Act Taxonomy

The 'Meeting Recorder Dialogue Act' (MRDA) taxonomy was developed and used to annotate dialogue acts in the International Computer Science Institute (ICSI) Meeting Corpus (Alexandersson et al., 2005). Labels in this taxonomy consist of a GENERAL TAG that may be followed by one or more SPECIAL TAGS and an optional DISRUPTION MARK. The general tags consist of:

1. Statement;
2. Question;
3. Backchannel; and
4. Floor Management.

Augmenting these general tags are forty special tags that describe aspects like positive, negative or uncertain response, restatements, politeness mechanisms and other functions. Disruptions forms are also marked by labels such as 'interrupted by other speaker', 'abandoned by speaker', and 'indecipherable'.

In their description of annotating the ICSI corpus, however, Alexandersson et al. (2005) noted that only 82 labels occur more than 100 times. The vast majority of the total 2050 labels occur very infrequently, with almost 1000 tags occurring once across the entire dataset. This strongly suggests that the MRDA taxonomy is too fine-grained to be practically useful for classification purposes. MRDA's top-level general tags also fail to accommodate speech acts such as the commissive acts that are required for our work.

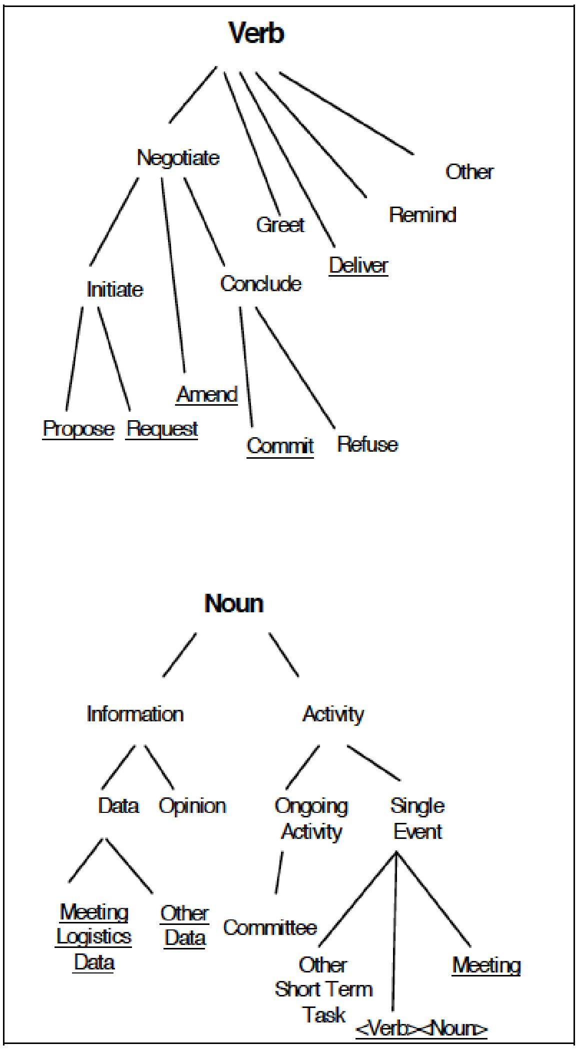


FIGURE 2.2: The Cohen, Carvalho and Mitchell (CCM) taxonomy of email speech acts. The underlined category labels represent classes for which they built classifiers.

Finally, and more fundamentally, the MRDA taxonomy is focused on speech annotation rather than text annotation; many of the special tags such as 'interrupted by other speaker' and 'abandoned by speaker' are generally not applicable to asynchronous, message-based email conversation.

2.3.8 Cohen, Carvalho and Mitchell's Taxonomy

Cohen, Carvalho, and Mitchell (2004) approached the classification of speech acts in email by developing separate taxonomies of verbs and nouns to define a taxonomy of message-level EMAIL ACTS, as shown in Figure 2.2. We refer to this as the Cohen, Carvalho and Mitchell (CCM) Taxonomy. Their verb taxonomy includes a number of categories that cater for request and commitment content: Request, Commit, Propose, Amend and Refuse. The nouns cater for common entities such as meetings, but due to the need to cater for open domain email data, also includes categories such as Other Data and Other Short Term Task.

Unfortunately, while their definitions fit their corpus of email data well, the CCM taxonomy mixes concepts of conversation state—whether an email message initiates, continues or concludes an email conversation—and illocutionary force in a manner that ties the definition of specific speech acts to particular conversation states. Conversation state is modelled through abstract categories such as Initiate, Conclude, and Negotiate, as shown in Figure 2.2. The presence of these conversation states in the taxonomy is problematic because, rather than consider conversation state and illocutionary force as orthogonal aspects of annotation, as in DAMSL or other dialogue act taxonomies, Cohen et al. define their Email Acts as children of these conversation states. As a result, every Commit act, for example, is defined as being part of the Conclude conversation state. Under this definition, Commit acts can only be marked in email messages that finish an email thread, which clearly is not a constraint that can cater to the range of real-world email communication.

Defining the speech act categories in terms of specific conversation states in this way also requires the creation of separate, overlapping categories in the taxonomy: the Propose and Amend acts, for example, differ only on whether the act occurs in a message that is the first in a new email thread or in a message that continues an ongoing email thread. These differences do not relate to the illocutionary force of the speech act being performed, but rather to the conversational context in which they occur. This mixing of concerns also causes problems for the coverage of the definitions. It is unclear, for example, how a commitment such as:

(2.22) *I'll send you the document with further details*

that occurs in a conversation-initiating email should be classified, since no defined class caters for this particular combination of a commitment speech act occurring in the first message of a thread.

While acknowledging that a single email message may contain multiple acts, the CCM taxonomy is also premised around avoiding the labelling of multiple acts for a single message. Partly because of this, the taxonomy does not cleanly capture the relationships between categories. It is not at all apparent, for example, that each of the Propose, Amend and Commit acts share the property of committing the sender. Partly this is due to their organising the taxonomy by categories that capture conversation state, but the decision to create separate categories (Propose and Amend) for speech acts that have both a directive and commissive function also affects the clarity of the taxonomy. Representing the directive and commissive components independently (as types of Request or Commit acts respectively), and applying multiple categories to each message, would actually make the speech act structure clearer. It is also interesting to note that the inter-annotator agreement is lowest for the Commit and PROPOSE categories, quite possibly because of confusion between these overlapping categories that share illocutionary force but differ on the conversation state in which they occur.

Cohen et al. note that the decision to create additional categories that combine atomic speech acts in this way is motivated by a desire to "reflect observed linguistic behavior [rather] than to reflect any abstract view of the space of possible speech acts"(Cohen, Carvalho, and Mitchell, 2004, p. 310). We share similar goals of wanting to reflect the nature of requests and commitments in actual email in a taxonomy for requests and commitments in email (which is why we base our own taxonomy on the results of multiple annotation experiments performed with a real-world corpus of business email). We also, however, want to ensure, as much as possible, that our categorisation does not unnecessarily combine atomic linguistic phenomena, and thus cloud the interpretation and analysis of any application of the definitions.

Of most interest to our needs, the Carvalho and Cohen taxonomy defines both commit and request acts. Both are, however, defined in terms of a static component of a conversation; requests are considered to be INITIATE acts, while commitments are defined as CONCLUDE acts. Additionally, the lack of detail in the published definitions leaves some utterances without a clear classification. For example, it is unclear whether an act of advice or suggestion such as:

(2.23) *I think you should include the figures in section 2*

should be classified as a REQUEST. Similarly, no guidance is given for how to classify conditional requests and commitments, except for a specific class of acts that request a response and conditionally commit the sender if the recipient responds (e.g., *let's do lunch*). These are identified as PROPOSE acts.

Overall, while there is strong alignment with our interest in requests and commitments, the focus on both propositional content, in the form of their noun ontology, and the limitations that stem from integrating conversation state in the speech act definitions led us to not adopt the CCM taxonomy in our work.

2.3.9 Leuski's Taxonomy

Leuski (2004) offers yet another speech-act-inspired taxonomy and uses it to categorise email messages with the aim of distinguishing the roles of different email authors based on the patterns of speech act usage in their incoming and outgoing email. He uses labels at the message level to attempt to identify roles (for example, graduate student and research assistant) and to determine the relationship between a pair of interlocutors (for example, a *supervisor of* relationship between a research adviser and a graduate student).

Leuski's taxonomy focuses heavily on requests and includes the following categories:

1. REQUEST INFO: e.g., *Do you have the url?*
2. REQUEST ADVICE: e.g., *What should I do next?*
3. REQUEST ACTION: e.g., *Please reserve a room.*

4. Request Meeting: e.g., *Let (sic) meet and discuss this.*
5. Plan: e.g., *We are going to do ...*
6. Provide Info: e.g., *Here is the url you wanted.*

Leuski's chosen categories unfortunately are limited in their detail and scope. The published category definitions are limited to a single out-of-context example sentence or phrase for each category. These phrases are included in the set of categories which we enumerate above. It is unclear how other utterances, such as requests for permission, should be coded. As with Camino's taxonomy, the focus is solely on requests; there is no category to mark commitments, such as promises or offers.

The fairly nuanced distinctions adopted between the different classes of request speech acts identified also increases the potential for ambiguity in labelling, particularly given the lack of detailed definitions. It is unclear, for example, how the following utterance should be labelled:

(2.24) *Let me know if you'd like to meet on Friday.*

Should this utterance in Example (2.24) be labelled as a Request Info because the sender expects an answer, or should it be a Request Action because the sender would like the recipient to organise a meeting, or should it be a Request Meeting? In Leuski's work, it is understandable that such distinctions may be salient for distinguishing between roles (e.g., a manager might request action from their subordinates, while the subordinate might request advice in return). For our work, however, the value derived from such distinctions does not justify the additional complexity and ambiguity that arise in attempting to identify the boundaries between these classes.

Together, these issues make Leuski's categories unable to be easily adopted for robust annotation or computational classification of requests and commitments. We do, however, draw insight from Leuski's idea of exploring patterns of speech acts over time between different interlocutors, and consider this to be an interesting application that could be built on top of our own analysis, classifiers and email platform.

2.3.10 The SmartMail Taxonomy

The SmartMail system (Corston-Oliver et al., 2004) attempts to automatically extract and reformulate action items from email messages for the purpose of adding them to a user's to-do list. Unlike most of the other taxonomies discussed, which focus on categorising email at the message level, the SmartMail taxonomy was designed to be applied at the sentence level.

The taxonomy of SmartMail categories is actually a mixture of syntactic and functional categories. The set of available categories are:

- Salutation;
- Chit-chat: social discussion unrelated to the main purpose of the email message;
- Task: items that can be added to an ongoing to-do list;

- Meeting: a proposal to meet;
- Promise;
- Farewell;
- Sig Name: the name component of an email signature;
- Sig Affiliation: the affiliation component of an email signature;
- Sig Title: the title component of an email signature;
- Sig Location: the location or address component of an email signature;
- Sig Phone: the phone number component of an email signature;
- Sig Email: the email address component of an email signature;
- Sig URL: the URL component of an email signature;
- Sig Other: any other components of an email signature; and
- none of the above: A default category to ensure every utterance can be labelled.

For requests and commitments, the three categories of interest are: Task, Meeting and Promise.

Unfortunately, in Corston-Oliver et al.'s published work, no mention is made of the use of the promise category, and no explanation is given for what constitutes a promise. Meeting speech acts are defined as "a proposal to meet", without any further detail. Tasks are defined as sentences that "look like an appropriate item to add to an ongoing 'to do' list", with the explicit exclusion of simple factual questions on the basis that the act of responding fulfils any associated obligation (meaning nothing is placed on a task list).

What constitutes a "simple factual question" is not specified, and it is not clear how to distinguish such requests from those that would result in a new action being added to a task list. Consider, for example, the utterance:

(2.25) *When is the contract expiring?*

This utterance might be answered immediately without the need to add anything to a task list if the recipient knows the response. Alternatively, it might require the recipient to issue requests to other people or systems; tasks that might reasonably be added to the recipient's task list. In either case, an obligation is placed on the recipient, and unlike Corston-Oliver et al., we believe this should be reflected in the classification of the utterance as a request under either interpretation.

Corston-Oliver et al.'s annotators were also restricted to applying a single tag to each sentence, meaning that a sentence could not embody both a request (task) and a commitment (promise), which, as we illustrated earlier in Example (2.26), is an artificial restriction.

Overall, the SmartMail taxonomy is close to what we adopt for our work, in that we use a single category for requests and another category for commitments. This similarity, however, comes with a significant caveat: that we are much more careful and methodical in defining which utterances are and are not requests or commitments.

2.3.11 Bennett and Carbonell's Taxonomy

Bennett and Carbonell (2005) explored the automatic classification of ACTION ITEMS in email. They define an action item as "an explicit request for information that requires the recipient's attention or a required action", and asked annotators to label both messages and segments of continuous text or sentences containing action items. They report agreements of κ=0.85 at the message level and κ=0.82 at the sentence level for recognising action items. Interestingly, they note that most disagreements were due to "different interpretations of conditional statements". Unfortunately, their definition of action items is quite restrictive, excluding all implicit requests. Their definition also requires value or importance judgements to be made about requests, which we seek to avoid due to the inherently context-dependent and subjective nature of such judgements. Bennett and Carbonell (2005) cite the following utterance as an example of an action item: *If you would like to keep your job, come to tomorrow's meeting.* By way of contrast, they cite the following utterance as a non-action item: *If you would like to join the football betting pool, come to tomorrow's meeting.* Unlike Bennett and Carbonell, we consider both these utterances to be (conditional) requests.

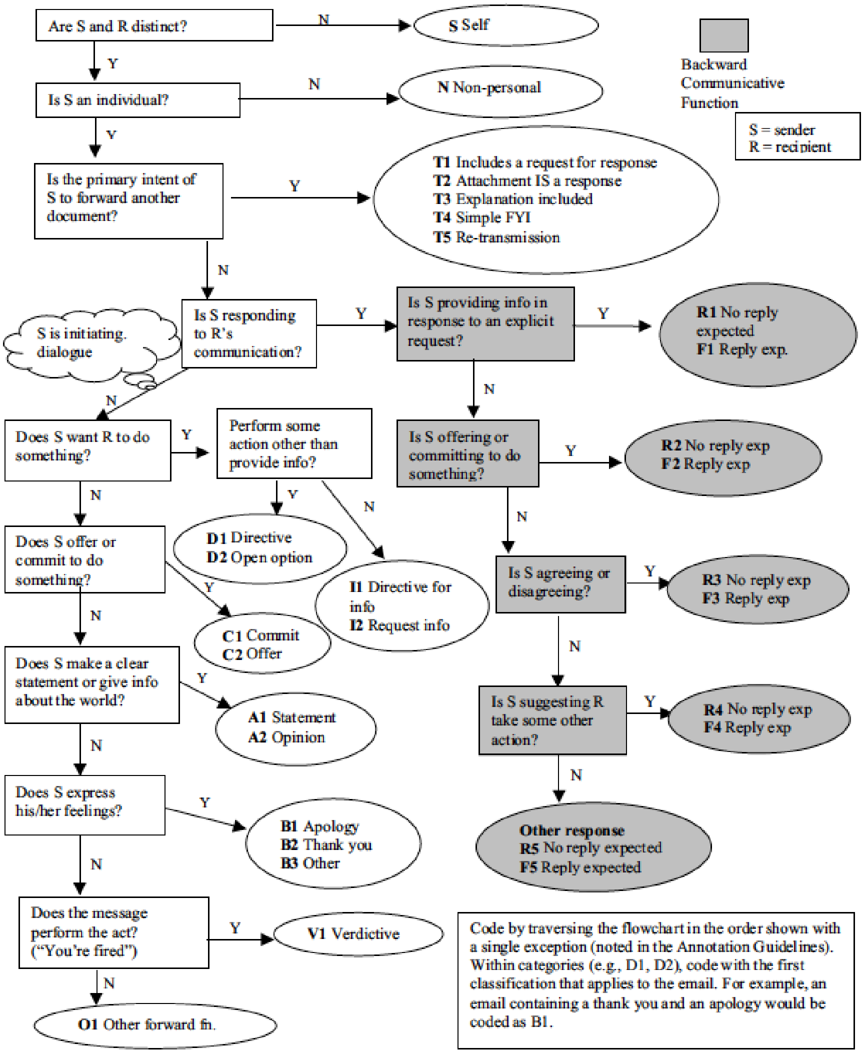
Ultimately, we adopt an approach that is similar to Bennett and Carbonell's in that we create definitions, albeit more comprehensive ones, that define our phenomena of interest (requests and commitments) and label these at both the sentence and message-level across our corpus. In terms of specifics, however, Bennett and Carbonell's action item definition is too narrow and lacking in detail to be directly applied. Additionally, as with several of the other taxonomies surveyed, commitments are not accounted for, despite their correlation with action items that occur in outgoing email.

2.3.12 Goldstein and Sabin's Taxonomy

In their work on categorising email messages and identifying different email genres, Goldstein and Sabin (2006) defined their own message-level taxonomy of email speech acts that includes twenty-three speech act categories. This taxonomy is shown in Figure 2.3, in the form of an annotation decision tree.

Goldstein and Sabin's taxonomy imports the concepts of forward looking and backward looking functions from DAMSL. Rather than keeping these aspects separate, as they are in DAMSL, Goldstein and Sabin choose to couple specific conversation states within their speech act definitions. Like the CCM taxonomy, this results in many of the speech act categories in the Goldstein and Sabin taxonomy being distinguished only by their sequence in conversation, rather than by the illocutionary force of the utterance. For example, request and commitment speech acts that respond to previous acts are classified into different categories depending on whether further response is expected. This distinction is, however, not captured for requests or commitments in an email message that initiates an email conversation.

Another limiting factor is that requests and commitments are defined as being mutually exclusive; under Goldstein and Sabin's application rules, a message cannot simultaneously request something from the recipient and commit the message sender.


Are S and R distinct?
N
S Self
Backward Communicative Function
Y
Is S an individual?
N
N Non-personal
S = sender
R = recipient
Y
Is the primary intent of S to forward another document?
Y
T1 Includes a request for response
T2 Attachment IS a response
T3 Explanation included
T4 Simple FYI
T5 Re-transmission
N
S is initiating. dialogue
Is S responding to R's communication?
Y
Is S providing info in response to an explicit request?
Y
R1 No reply expected
F1 Reply exp.
N
N
Does S want R to do something?
Y
Perform some action other than provide info?
Is S offering or committing to do something?
Y
R2 No reply exp
F2 Reply exp
N
Y
N
D1 Directive
D2 Open option
N
Does S offer or commit to do something?
Y
I1 Directive for info
I2 Request info
Is S agreeing or disagreeing?
Y
R3 No reply exp
F3 Reply exp
N
C1 Commit
C2 Offer
N
Does S make a clear statement or give info about the world?
Y
A1 Statement
A2 Opinion
Is S suggesting R take some other action?
Y
R4 No reply exp
F4 Reply exp
N
N
Does S express his/her feelings?
Y
B1 Apology
B2 Thank you
B3 Other
Other response
R5 No reply expected
F5 Reply expected
N
Does the message perform the act? ("You're fired")
Y
V1 Verdictive
N
O1 Other forward fn.

Code by traversing the flowchart in the order shown with a single exception (noted in the Annotation Guidelines). Within categories (e.g., D1, D2), code with the first classification that applies to the email. For example, an email containing a thank you and an apology would be coded as B1.

FIGURE 2.3: The Goldstein and Sabin taxonomy of email speech acts, including a decision tree of definitions (Goldstein and Sabin, 2006, p. 3).

This limitation is problematic for both message and utterance-level annotation. Consider an utterance such as

(2.26) *Let me know if you'd like a copy of the document*

This utterance requests a response and conditionally commits the sender to sending the document. Under Goldstein and Sabin's taxonomy it is quickly apparent that there is no valid classification for this utterance; it is only possible to label a Commit or Offer if the annotator answers 'No' to the question 'Does S want R to do something?'. Our annotation experiments show that such utterances are relatively frequent in business email. Messages that convey both commitments and requests are even more frequent.

To their credit, Goldstein and Sabin include conditionality in their definition of commitments in that separate categories are defined for Offer and Commit acts, depending on whether the commitment is conditional. These categories are, however, defined only by example phrases such as *Would you like me to ...* and *I can ...* . As with Khosravi and Wilks' taxonomy, such form-based definitions are problematic, for the reasons outlined in the introduction to Section 2.3, there is not a simple mapping between linguistic forms and the illocutionary force of utterances.

Overall, the complexity of Goldstein and Sabin's taxonomy is not justified by our requirements, and some of the category definitions lack the detail required for the unambiguous annotation of real-world data. Additionally, as with the CCM taxonomy, the limitations that stem from mixing conversational state in the speech act definitions discouraged us from adopting this categorisation.

2.3.13 A Summary of Speech Act Taxonomies

In his review of available speech act taxonomies, Verschueren makes the point that:

> "There is no logical end to the number of possibilities (of speech act taxonomies); and every new taxonomy will necessarily prevent us from seeing certain things, because each one focuses on a selection from a wide range of relevant dimensions of variation."
> (Verschueren, 1983, p. 173)

As we discuss above in our analysis of each categorisation, none of the speech act taxonomies we have surveyed has quite the right focus on relevant dimensions to make it sufficient for our purposes.

As we have seen, some categorisations, such as the VerbMobil taxonomy, are limited to their application within a particular domain; many others, including the SmartMail, Leuski and CCM taxonomies, lack detailed definitions that would enable an unambiguous classification at the utterance level. Still others, including the CCM and Goldstein and Sabin's taxonomy, intertwine conversation state into their speech act categories, which results in an explosion of possible categories, not all of which are represented in the taxonomy, even when those categories have been empirically observed.

Additionally, despite the importance of commitments, it is notable from our survey how little research has specifically focused on understanding commitments, particularly in email. Waldvogel (2002) is one exception who has qualitatively examined commitment in workplace email from a speech act perspective. In a small corpus of email, Waldvogel found commissive acts made up 6% of speech act functions and were the main function of only 4% of email messages.

Our research suggests that commitments have a much bigger role to play; across our real-world email corpus, we have found somewhere between 22% and 37% of messages were agreed by annotators to contain commitments. In our fine-grained annotation experiments, at least one annotator marked a commitment in 48% of the email messages. Suffice to say that, at least in the data we have worked with, commitments occur more frequently than the previous studies have suggested. All of the taxonomies above each have deficiencies in dealing with commitments. Some, such as the Leuski and Camino taxonomies, do not deal with commitments at all. Others, such as the SmartMail taxonomy, do not cater for conditional commitments.

As a result of this collection of issues, we do not adopt any single existing taxonomy or set of definitions for our work in this thesis. Instead, as noted in our discussion of each taxonomy, we draw on a range of aspects and insights from the categories and approaches that are employed in each of the taxonomies we have surveyed to guide our own definitions of request and commitment acts that we elaborate on in Chapter 3.

2.4 A Review of Actionable Email Systems

Building on the analyses and categorisations of speech acts surveyed above, there are also a number of systems and ideas that have been prototyped to improve the way that actionable, task-oriented content is dealt with in email software.

Within the literature of systems that have been developed to assist users work with actionable content in email messages, there are two main strands of research: manual systems that provide the opportunity for users to identify tasks within their email content, and automated systems that focus on creating fully or semi-automatic techniques for identifying actionable content. Below we review a range of significant systems in the literature from which we look to draw insights and lessons for our own research.

2.4.1 Systems Supporting Manual Task Identification

In order to ease the burden on users, several systems have been proposed which increase the visibility and role of email threads or conversations within email clients, relative to commonly used commercial email applications. The aim of these systems is generally to make better use of the conversational structure in email. In this section, we step through some of the major systems that have promoted manual task identification and management.

Some, such as the Coordinator system, focus on patterns of acts over conversations, while others, such as the TaskMaster and ReMail systems, focus more on interface issues around how to keep tasks visible to encourage user action.

A common limitation of systems that require manual task identification is that their adoption by users is frequently hampered by the effort and overhead that people need to expend in manually identifying tasks. Rather than focus on these limitations, however, our aim is to learn from these systems about ways to intuitively integrate tasks into existing email interfaces, as well as positive and negative design issues that promoted or hindered user adoption.

2.4.1.1 The Coordinator

Winograd and Flores (1986) were pioneers in considering electronic messaging from a theoretical linguistic point of view, and their ideas have been highly influential. Their

Coordinator system drew on Speech Act Theory to model conversations within organisations as networks of obligation acts. Users of the Coordinator system were required to select explicitly among a small set of predefined speech acts in order to specify the commissive and directive acts embodied in each message they sent. Users were also required to identify relationships with other speech acts and important temporal characteristics that might be used to identify when communication breakdowns had occurred, that is, when a conversation that should have resulted in completion of some action had failed to reach a successful conclusion in the expected time frame. The idea was that this additional information from each sender would allow the status of conversations actioned via electronic messages to be easily monitored and reported.

Unfortunately, deployment of the Coordinator system largely failed in practice. According to several studies, e.g., (Carasik and Grantham, 1988; Robinson, 1990), this was primarily because the burden of manually classifying every message into the explicit categories that the Coordinator imposed proved too onerous, or at least not worth the effort for many users. Another significant and fundamental issue was that the rigid approach of the Coordinator, which obliged users to explicitly identify communicative acts in their messages, proved too restrictive for the highly flexible and nuanced realities of workplace conversation. People were reluctant to remove all ambiguity from the meaning of their messages by making their intentions so explicit. While such explicitness seems effective from a productivity perspective, Carasik and Grantham (1988) found that this requirement was in conflict with common communicative approaches like the use of indirect language to soften directive acts, and with the nature of human communication, which, by accident or design, permits and even encourages ambiguity which is resolved collaboratively between interlocutors.

The lessons from the design and application of the Coordinator system are numerous. High on that list is to ensure that any systems that attempt to create or identify structure within conversations should be highly flexible and adaptive, to deal with the needs of real workplace conversations. This is something that we focus carefully on in the application of our own work to email software in Chapter 7.

2.4.1.2 Issue Based Information System (IBIS)

A similar idea to the Coordinator system was the Issue Based Information System (IBIS) (Rittel and Kunz, 1970). Begeman and Conklin later developed gIBIS, a graphical implementation of IBIS for Sun workstations (Conklin and Begeman, 1988). gIBIS was a hypertext environment for the structured discussion of design issues. gIBIS used a stringent classification scheme to organise the data. In the structured graph-based visualisation offered, there are three node types (issues, positions, arguments) that identify the propositional type of a piece of content, and eight link types that capture the role it has relative to other content. The eight link types were:

1. Responds-to;
2. Questions;
3. Supports;
4. Objects-to;

5. Specializes;
6. Generalizes;
7. Refers-to; and
8. Replaces.

Similar to the issues identified with the Coordinator system, studies have observed that the overhead of explicitly specifying structure can be an obstacle to its use (Conklin and Begeman, 1988). The tendency of discussions to be sidetracked by debates on the correct use of the IBIS structure further suggests that an imposed structure may be too rigid.

Again, the lesson for our work from IBIS is that we should work to support rather than constrain people's communication, and, as much as possible, minimise the burden on people in making any changes to the processes that people follow in writing, and sending emails.

2.4.1.3 Zest

More recently, the Zest system (Yee, 2002) attempted to provide conversation-based visualisation of mailing list messages. Unfortunately, despite the author's criticism of the manual work required for users to classify messages in previous systems, the Zest system requires users to manually insert textual symbols called criticons, which mark paragraphs within emails with one of four types:

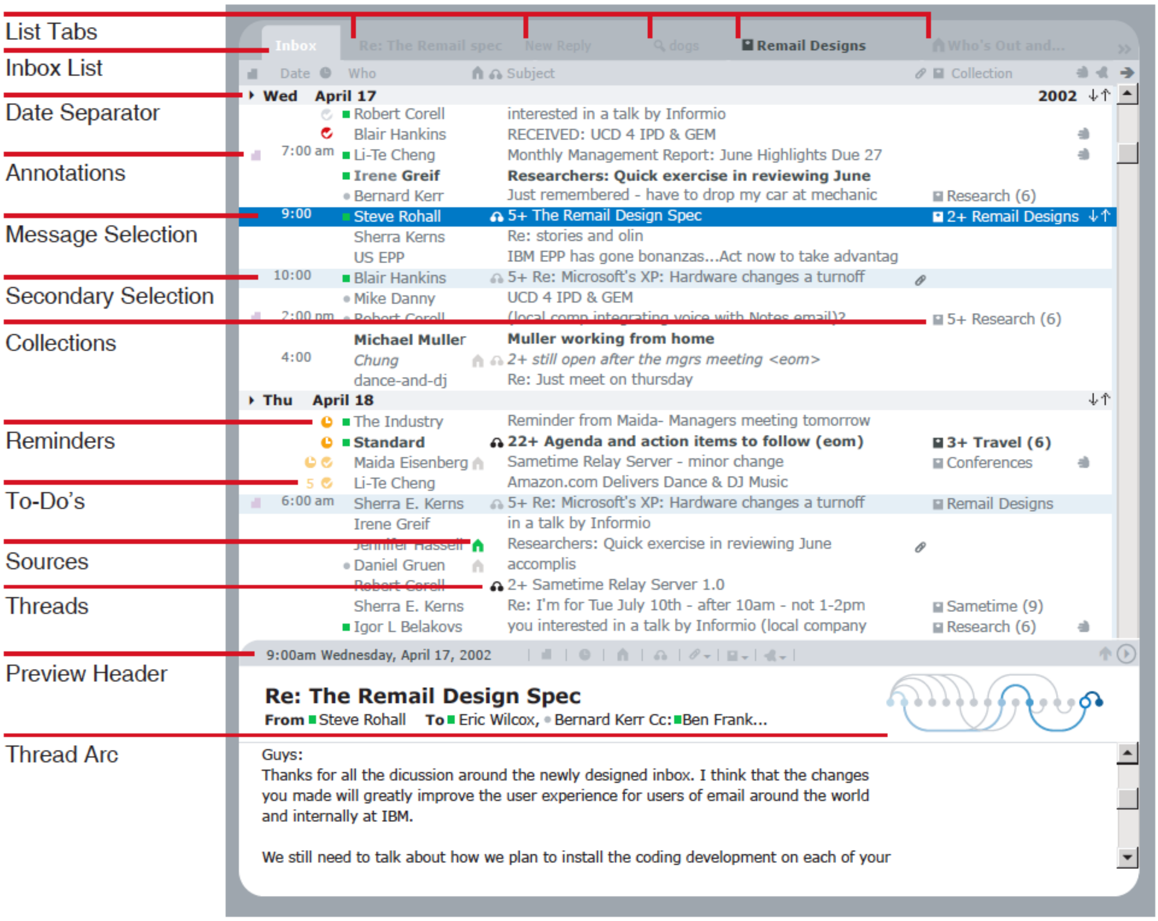
1. Question;
2. Statement;
3. Supporting Argument; or
4. Opposing argument.

This four-type categorisation was claimed to provide "most of the useful semantics of IBIS, but [be] simpler and easier to remember" (Yee, 2002, p. 124). Yee also includes a separate 'special criticon' to mark a section of text as a proposed resolution of a discussion and to flag an entire thread as resolved. In this way, Zest focuses on conversation state, and on making the structure of contributions within each thread explicit.

Zest aims to allow participants in a mailing list to see quickly which issues are resolved or open, and to determine which questions have already been asked and answered. As with the gIBIS and Coordinator systems, this benefit only comes after investing a significant amount of time tagging the messages manually. Our aim is to alleviate some of this burden through the development and application of automated and semi-automated classification techniques.

2.4.1.4 ReMail

The Remail system (Rohall and Gruen, 2002; Kerr and Wilcox, 2004) took a design-focused approach to changing the way users interact with email. One aspect of this was



Date	Who	Subject	Collection
Wed April 17			2002
	Robert Corell	interested in a talk by Informio	
	Blair Hankins	RECEIVED: UCD 4 IPD & GEM	
7:00 am	Li-Te Cheng	Monthly Management Report: June Highlights Due 27	
	Irene Greif	**Researchers: Quick exercise in reviewing June**	
	Bernard Kerr	Just remembered - have to drop my car at mechanic	Research (6)
9:00	Steve Rohall	5+ The Remail Design Spec	2+ Remail Designs
	Sherra Kerns	Re: stories and olin	
	US EPP	IBM EPP has gone bonanzas...Act now to take advantag	
10:00	Blair Hankins	5+ Re: Microsoft's XP: Hardware changes a turnoff	
	Mike Danny	UCD 4 IPD & GEM	
2:00 pm	Robert Corell	(local comp integrating voice with Notes email)?	5+ Research (6)
	Michael Muller	**Muller working from home**	
4:00	*Chung*	*2+ still open after the mgrs meeting <eom>*	
	dance-and-dj	Re: Just meet on thursday	
Thu April 18			
	The Industry	Reminder from Maida- Managers meeting tomorrow	
	Standard	**22+ Agenda and action items to follow (eom)**	**3+ Travel (6)**
	Maida Eisenberg	Sametime Relay Server - minor change	Conferences
5	Li-Te Cheng	Amazon.com Delivers Dance & DJ Music	
6:00 am	Sherra E. Kerns	5+ Re: Microsoft's XP: Hardware changes a turnoff	Remail Designs
	Irene Greif	in a talk by Informio	
	Jennifer Hassell	Researchers: Quick exercise in reviewing June	
	Daniel Gruen	accomplis	
	Robert Corell	2+ Sametime Relay Server 1.0	
	Sherra E. Kerns	Re: I'm for Tue July 10th - after 10am - not 1-2pm	Sametime (9)
	Igor L Belakovs	you interested in a talk by Informio (local company	Research (6)

9:00am Wednesday, April 17, 2002

Re: The Remail Design Spec
From Steve Rohall **To** Eric Wilcox, Bernard Kerr Cc: Ben Frank...

Guys:
Thanks for all the dicussion around the newly designed inbox. I think that the changes you made will greatly improve the user experience for users of email around the world and internally at IBM.

We still need to talk about how we plan to install the coding development on each of your

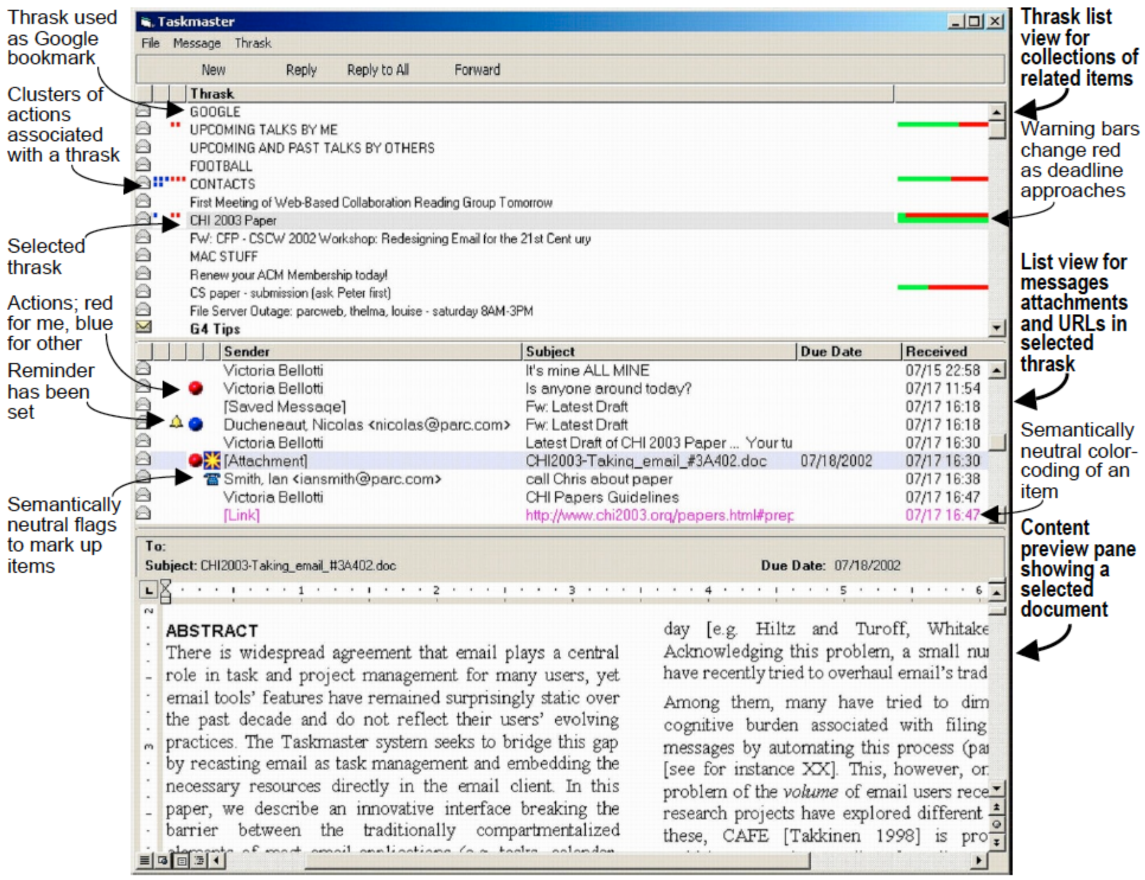
Figure 2.4: The ReMail system.

their approach to dealing with actionable items in email. Users were able to manually mark messages with a to-do label, which would then associate the message with a to-do List. Once a to-do item was completed, users could then mark it as a Completed to-do item, and the message would be archived with other completed tasks.

To-do items could be accessed from the Calendar and Inbox, and the top message in the to-do list would resurface each day in the inbox to keep outstanding tasks in sight, as can be seen in Figure 2.4. This resurfacing was explicitly designed to address tasks in the inbox scrolling out of sight under the deluge of incoming mail. While ReMail introduced a range of interesting and novel ideas for improving email, their focus on task management was minimal. Their approach deals only with tasks at the message level, and relies entirely on user input to identify and manage tasks.

2.4.1.5 TaskMaster

The Taskmaster system (Bellotti et al., 2003) introduces the concept of thrasks, which are threaded, task-centric collections of messages, drafts, documents and URL


Thrask used as Google bookmark
Clusters of actions associated with a thrask
Selected thrask
Actions; red for me, blue for other
Reminder has been set
Semantically neutral flags to mark up items
Thrask list view for collections of related items
Warning bars change red as deadline approaches
List view for messages attachments and URLs in selected thrask
Semantically neutral color-coding of an item
Content preview pane showing a selected document
Taskmaster
File
Message
Thrask
New
Reply
Reply to All
Forward

Thrask
GOOGLE
UPCOMING TALKS BY ME
UPCOMING AND PAST TALKS BY OTHERS
FOOTBALL
CONTACTS
First Meeting of Web-Based Collaboration Reading Group Tomorrow
CHI 2003 Paper
FW: CFP - CSCW 2002 Workshop: Redesigning Email for the 21st Cent ury
MAC STUFF
Renew your ACM Membership today!
CS paper - submission (ask Peter first)
File Server Outage: parcweb, thelma, louise - saturday 8AM-3PM
G4 Tips

Sender	Subject	Due Date	Received
Victoria Bellotti	It's mine ALL MINE		07/15 22:58
Victoria Bellotti	Is anyone around today?		07/17 11:54
[Saved Message]	Fw: Latest Draft		07/17 16:18
Ducheneaut, Nicolas <nicolas@parc.com>	Fw: Latest Draft		07/17 16:18
Victoria Bellotti	Latest Draft of CHI 2003 Paper ... Your tu		07/17 16:30
[Attachment]	CHI2003-Takinq_email_#3A402.doc	07/18/2002	07/17 16:30
Smith, Ian <iansmith@parc.com>	call Chris about paper		07/17 16:38
Victoria Bellotti	CHI Papers Guidelines		07/17 16:47
[Link]	http://www.chi2003.orq/papers.html#prep		07/17 16:47

To:
Subject: CHI2003-Taking_email_#3A402.doc **Due Date:** 07/18/2002

ABSTRACT
There is widespread agreement that email plays a central role in task and project management for many users, yet email tools' features have remained surprisingly static over the past decade and do not reflect their users' evolving practices. The Taskmaster system seeks to bridge this gap by recasting email as task management and embedding the necessary resources directly in the email client. In this paper, we describe an innovative interface breaking the barrier between the traditionally compartmentalized

day [e.g. Hiltz and Turoff, Whitake
Acknowledging this problem, a small nu
have recently tried to overhaul email's trad

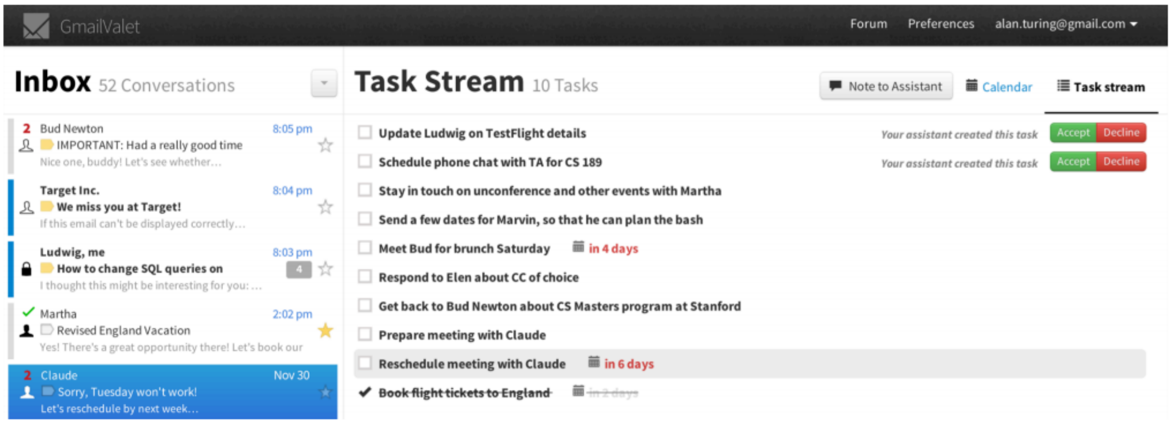
Among them, many have tried to dim
cognitive burden associated with filing
messages by automating this process (pa
[see for instance XX]. This, however, or
problem of the *volume* of email users rece
research projects have explored different
these, CAFE [Takkinen 1998] is pro

FIGURE 2.5: The TaskMaster system. The top pane provides the thrask list, the middle pane shows the message and other thrask member items, and the bottom pane provides a preview of selected content.

links. The Taskmaster interface is shown in Figure 2.5. Thrasks are created semi-automatically based on email threads in the inbox. Users are also able to add and remove messages to and from existing thrasks, or to create new thrasks manually, and can annotate thrasks and sub-items manually with meta-information about deadlines, reminders and actions.

In real-world use, it became apparent that Taskmaster requires constant user actions in order to maintain a useful set of thrasks. One of the small set of Taskmaster test subjects found that unread messages in Taskmaster quickly built up, making it too time consuming for him to recover after one week (Bellotti et al., 2003).

This process of dealing with a backlog of messages is often referred to as email triage. Specifically, email triage is the process of going through unhandled email and deciding what to do with it. Email triage can quickly become a serious problem for users as the amount of unhandled email grows. Neustaedter, Brush, and Smith (2005)



Inbox 52 Conversations

- 2 Bud Newton — 8:05 pm — IMPORTANT: Had a really good time — Nice one, buddy! Let's see whether...
- Target Inc. — 8:04 pm — We miss you at Target! — If this email can't be displayed correctly...
- Ludwig, me — 8:03 pm — How to change SQL queries on — 4 — I thought this might be interesting for you: ...
- ✓ Martha — 2:02 pm — Revised England Vacation — Yes! There's a great opportunity there! Let's book our
- 2 Claude — Nov 30 — Sorry, Tuesday won't work! — Let's reschedule by next week...

Task Stream 10 Tasks

- □ Update Ludwig on TestFlight details — *Your assistant created this task*
- □ Schedule phone chat with TA for CS 189 — *Your assistant created this task*
- □ Stay in touch on unconference and other events with Martha
- □ Send a few dates for Marvin, so that he can plan the bash
- □ Meet Bud for brunch Saturday — in 4 days
- □ Respond to Elen about CC of choice
- □ Get back to Bud Newton about CS Masters program at Stanford
- □ Prepare meeting with Claude
- □ Reschedule meeting with Claude — in 6 days
- ✓ ~~Book flight tickets to England~~ — ~~in 2 days~~

Figure 2.6: The Email Valet system.

investigated the problem of email triage by presenting interview and survey results that articulate user needs. Their results suggest the need for email user interfaces to provide additional salient information in order to bring important emails to the forefront against the background noise of other email messages. The inability of a user to keep using the Taskmaster system after one week of email build-up raises some doubt about the ability of the Taskmaster design to scale and function well in the real world, where email triage is all too common.

2.4.1.6 The Email Valet

The Email Valet system (Kokkalis et al., 2013) offers a novel twist on automated systems for identifying task-related content in email.[2] Rather than attempt to automatically extract tasks, the Email Valet offers an email client, shown in Figure 2.6, that interfaces with remote human assistants who have been recruited from a crowdsourcing marketplace to manually annotate each email with its implied tasks. The actions of the human assistants populate a task list that contains tasks extracted from the user's email messages.

In this way, the EmailValet system offers a novel hybrid of manual and automatic task detection: to the email user, it appears that tasks are automatically extracted, since they need expend no effort themselves. Behind the scenes, however, there are no algorithms or rules, but other people manually identifying tasks on behalf of the email user.

The Email Valet is a refreshingly innovative approach to dealing with accurate task detection in email. Kokkalis et al. note, however, that there are unresolved issues of security and privacy with their system, given the level of trust an email user must place

[2] Note that this system is unrelated to the EmailValet System developed by Macskassy, Dayanik, and Hirsh (1999), which was designed to detect urgent and important messages.

in anonymous human assistants. They also point at issues around the lack of context for the assistants extracting tasks, and the variable quality of the work done by those assistants, as potential barriers to adoption.

2.4.2 Systems Supporting Automated Task Identification

There is some work in the literature about extending the ideas of Winograd and Flores in office environments to remove much of the burden of manual classification that impeded uptake and use of systems like the Coordinator, TaskMaster and IBIS, as discussed above.

The possibilities of performing automatic task identification are appealing, though the challenges are significant, for all of the reasons that we discussed in Section 2.3. In this section, we review systems that have attempted to apply rule-based and statistical machine learning techniques to recognise requests, and sometimes commitments, in email messages. In many cases, the focus is on the algorithms themselves, rather than on how such techniques would be integrated into email software.

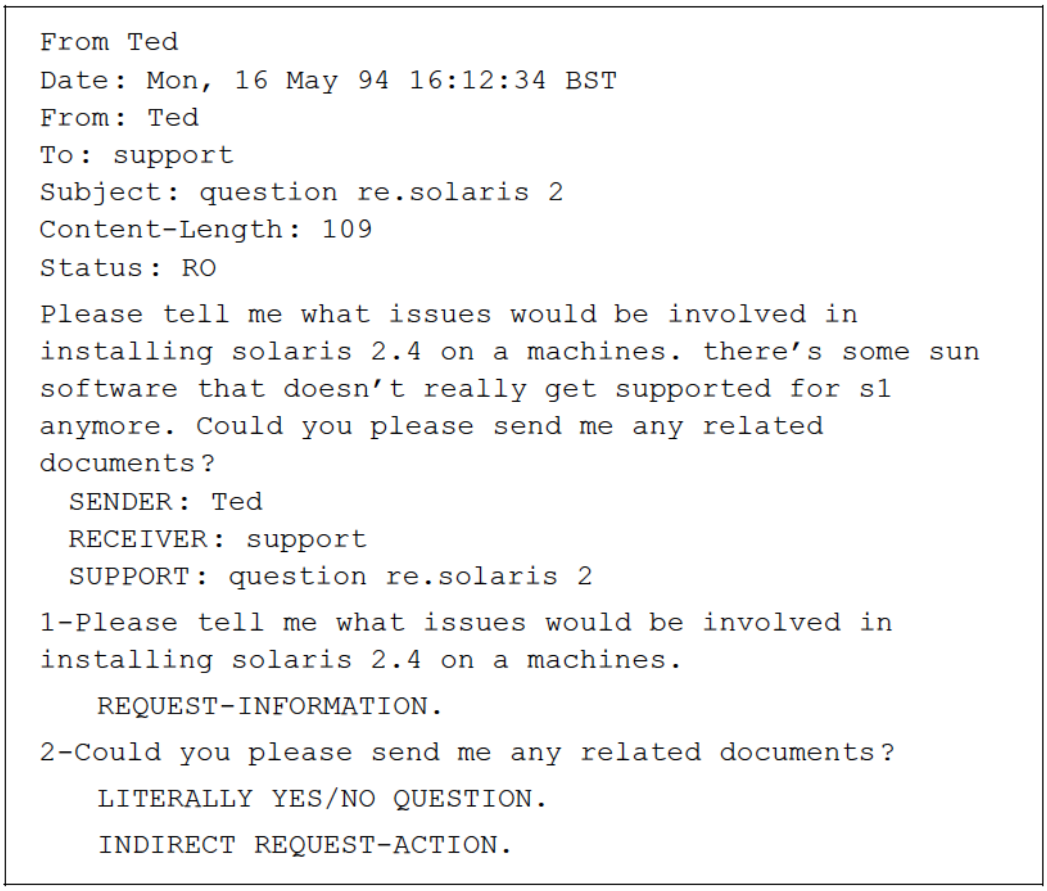
We look to gather promising approaches and insights from this work to guide our own statistical speech act classification work in Chapters 4 and 5.

2.4.2.1 Pyam

Khosravi and Wilks (Khosravi and Wilks, 1999) were among the first to automate message-level speech act classification in email for their Pyam system that they described as ‘Routing Email by Purpose Not Topic’. They adopted cue-phrase based rules to classify three classes of requests, Request-Action, Request-Information and Request-Permission, as described above in Section 2.3.5. The Pyam system would then extract the sentences that were deemed to represent requests of the above types, and provide a textual summary of the original message, as shown in Figure 2.7.

The Pyam prototype presents the users with labels from the full set of syntactically focused request categories that we discussed in Section 2.3.5. The distinctions within this set of categories might well have a role under the hood of Pyam, since it uses simple syntactic rules to identify which sentences should be marked with which category. Users, however, are highly unlikely to react differently based on whether a question arrives in a *yes/no* form or a *what/where/when* form. Such distinctions do not represent differences in illocutionary force, but in the surface form used to convey the sender's illocutionary force.

The corpus used by Pyam consisted of 1000 email messages from a university computer helpdesk. Rather than rely on *a priori* intuitions to hand-craft the cue phrases used to recognise requests, Khosravi manually annotated each sentence in the corpus, then used n-gram word frequencies for each class of request to build rules to recognise them. While this approach is generally applicable, the rules they extracted naturally reflect the specific domain of their email data, and are very specific to the computer support domain from which their email data was drawn. A dramatic drop in performance was observed when the system was trialled on data from a different domain.



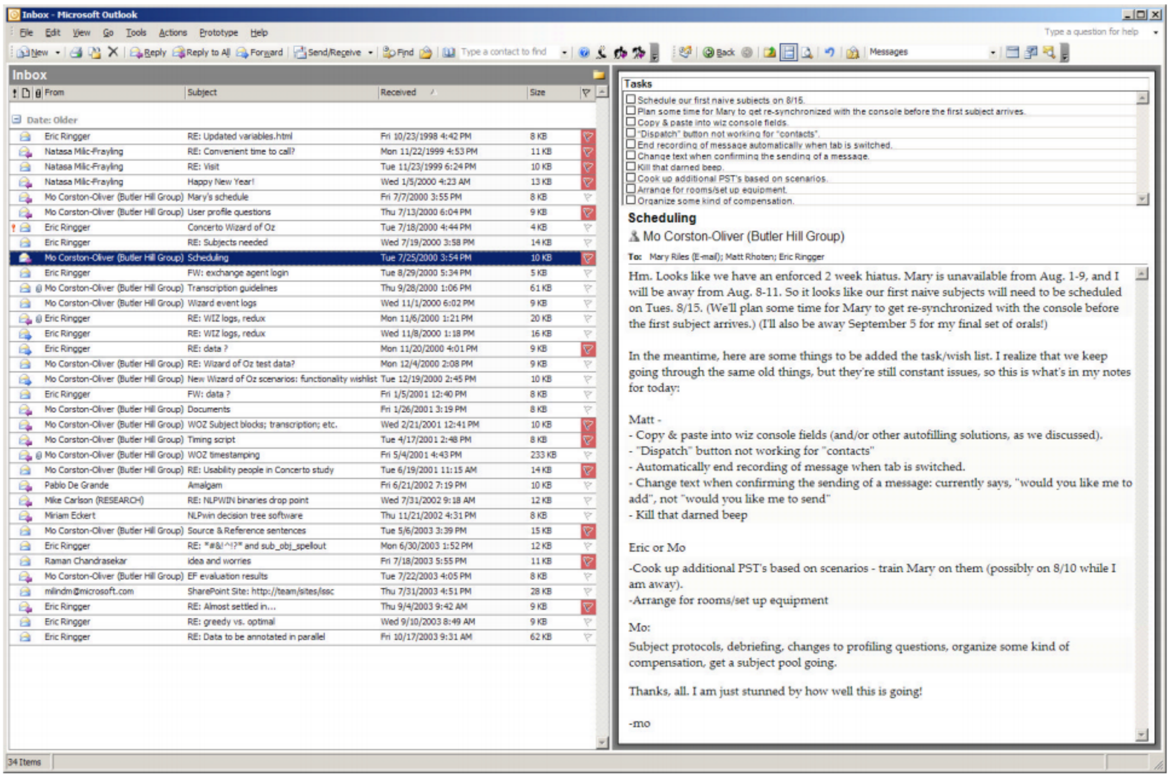
```
From Ted
Date: Mon, 16 May 94 16:12:34 BST
From: Ted
To: support
Subject: question re.solaris 2
Content-Length: 109
Status: RO
Please tell me what issues would be involved in
installing solaris 2.4 on a machines. there's some sun
software that doesn't really get supported for s1
anymore. Could you please send me any related
documents?
  SENDER: Ted
  RECEIVER: support
  SUPPORT: question re.solaris 2
1-Please tell me what issues would be involved in
installing solaris 2.4 on a machines.
    REQUEST-INFORMATION.
2-Could you please send me any related documents?
    LITERALLY YES/NO QUESTION.
    INDIRECT REQUEST-ACTION.
```

Figure 2.7: The Pyam email request system.

2.4.2.2 Ciranda

A more recent example is the work by Carvalho and colleagues which involved applying statistical text classification techniques to identifying 'email speech acts' (Cohen, Carvalho, and Mitchell, 2004; Carvalho, 2005; Carvalho and Cohen, 2006). Their work explicitly draws inspiration directly from Flores and Winograd's work. Their original Ciranda system (Cohen, Carvalho, and Mitchell, 2004) used machine learning-based classifiers for recognising a number of speech acts in email text. They performed manual segmentation of email messages into different functional parts as pre-processing, but did not explore the contribution this made to the performance of their various speech act classifiers, as we do in Chapter 4. For identifying requests, they report peak F-score of 0.69 against a majority class baseline accuracy of approximately 66%.

Cohen, Carvalho and Mitchell found that unweighted bigrams were particularly useful features in their experiments, outperforming other features applied. They later applied a series of text normalisations and n-gram feature selection algorithms to improve performance (Carvalho and Cohen, 2006). We adopt similar normalisations in



Inbox - Microsoft Outlook
File Edit View Go Tools Actions Prototype Help
Type a question for help
New Reply Reply to All Forward Send/Receive Find Type a contact to find Back Messages
Inbox
Date: Older
34 Items

From	Subject	Received	Size
Eric Ringger	RE: Updated variables.html	Fri 10/23/1998 4:42 PM	8 KB
Natasa Milic-Frayling	RE: Convenient time to call?	Mon 11/22/1999 4:53 PM	11 KB
Natasa Milic-Frayling	RE: Visit	Tue 11/23/1999 6:24 PM	10 KB
Natasa Milic-Frayling	Happy New Year!	Wed 1/5/2000 4:23 AM	13 KB
Mo Corston-Oliver (Butler Hill Group)	Mary's schedule	Fri 7/7/2000 3:55 PM	8 KB
Mo Corston-Oliver (Butler Hill Group)	User profile questions	Thu 7/13/2000 6:04 PM	9 KB
Eric Ringger	Concerto Wizard of Oz	Tue 7/18/2000 4:44 PM	4 KB
Eric Ringger	RE: Subjects needed	Wed 7/19/2000 3:58 PM	14 KB
Mo Corston-Oliver (Butler Hill Group)	Scheduling	Tue 7/25/2000 3:54 PM	10 KB
Eric Ringger	FW: exchange agent login	Tue 8/29/2000 5:34 PM	5 KB
Mo Corston-Oliver (Butler Hill Group)	Transcription guidelines	Thu 9/28/2000 1:06 PM	61 KB
Mo Corston-Oliver (Butler Hill Group)	Wizard event logs	Wed 11/1/2000 6:02 PM	9 KB
Eric Ringger	RE: WIZ logs, redux	Mon 11/6/2000 1:21 PM	20 KB
Eric Ringger	RE: WIZ logs, redux	Wed 11/8/2000 1:18 PM	16 KB
Eric Ringger	RE: data ?	Mon 11/20/2000 4:01 PM	9 KB
Mo Corston-Oliver (Butler Hill Group)	RE: Wizard of Oz test data?	Mon 12/4/2000 2:08 PM	9 KB
Mo Corston-Oliver (Butler Hill Group)	New Wizard of Oz scenarios: functionality wishlist	Tue 12/19/2000 2:45 PM	10 KB
Eric Ringger	FW: data ?	Fri 1/5/2001 12:40 PM	8 KB
Mo Corston-Oliver (Butler Hill Group)	Documents	Fri 1/26/2001 3:19 PM	8 KB
Mo Corston-Oliver (Butler Hill Group)	WOZ Subject blocks; transcription; etc.	Wed 2/21/2001 12:41 PM	10 KB
Mo Corston-Oliver (Butler Hill Group)	Timing script	Tue 4/17/2001 2:48 PM	8 KB
Mo Corston-Oliver (Butler Hill Group)	WOZ timestamping	Fri 5/4/2001 4:43 PM	233 KB
Mo Corston-Oliver (Butler Hill Group)	RE: Usability people in Concerto study	Tue 6/19/2001 11:15 AM	14 KB
Pablo De Grande	Amalgam	Fri 6/21/2002 7:19 PM	10 KB
Mike Carlson (RESEARCH)	RE: NLPWIN binaries drop point	Wed 7/31/2002 9:18 AM	12 KB
Miriam Eckert	NLPwin decision tree software	Thu 11/21/2002 4:31 PM	8 KB
Mo Corston-Oliver (Butler Hill Group)	Source & Reference sentences	Tue 5/6/2003 3:39 PM	15 KB
Eric Ringger	RE: "#&!^!?" and sub_obj_spellout	Mon 6/30/2003 1:52 PM	12 KB
Raman Chandrasekar	idea and worries	Fri 7/18/2003 5:55 PM	11 KB
Mo Corston-Oliver (Butler Hill Group)	EF evaluation results	Tue 7/22/2003 4:05 PM	8 KB
milindm@microsoft.com	SharePoint Site: http://team/sites/lssc	Thu 7/31/2003 4:51 PM	28 KB
Eric Ringger	RE: Almost settled in...	Thu 9/4/2003 9:42 AM	9 KB
Eric Ringger	RE: greedy vs. optimal	Wed 9/10/2003 8:49 AM	9 KB
Eric Ringger	RE: Data to be annotated in parallel	Fri 10/17/2003 9:31 AM	62 KB

Tasks

- □ Schedule our first naive subjects on 8/15.
- □ Plan some time for Mary to get re-synchronized with the console before the first subject arrives.
- □ Copy & paste into wiz console fields.
- □ "Dispatch" button not working for "contacts".
- □ End recording of message automatically when tab is switched.
- □ Change text when confirming the sending of a message.
- □ Kill that darned beep.
- □ Cook up additional PST's based on scenarios.
- □ Arrange for rooms/set up equipment.
- □ Organize some kind of compensation.

Scheduling

Mo Corston-Oliver (Butler Hill Group)

To: Mary Riles (E-mail); Matt Rhoten; Eric Ringger

Hm. Looks like we have an enforced 2 week hiatus. Mary is unavailable from Aug. 1-9, and I will be away from Aug. 8-11. So it looks like our first naive subjects will need to be scheduled on Tues. 8/15. (We'll plan some time for Mary to get re-synchronized with the console before the first subject arrives.) (I'll also be away September 5 for my final set of orals!)

In the meantime, here are some things to be added the task/wish list. I realize that we keep going through the same old things, but they're still constant issues, so this is what's in my notes for today:

Matt -
- Copy & paste into wiz console fields (and/or other autofilling solutions, as we discussed).
- "Dispatch" button not working for "contacts"
- Automatically end recording of message when tab is switched.
- Change text when confirming the sending of a message: currently says, "would you like me to add", not "would you like me to send"
- Kill that darned beep

Eric or Mo

-Cook up additional PST's based on scenarios - train Mary on them (possibly on 8/10 while I am away).
-Arrange for rooms/set up equipment

Mo:

Subject protocols, debriefing, changes to profiling questions, organize some kind of compensation, get a subject pool going.

Thanks, all. I am just stunned by how well this is going!

-mo

FIGURE 2.8: The SmartMail system.

the message-level classification work that we present in Chapter 4.

Ciranda only deals with classification at the message level, without explicitly considering which specific utterances embody each specific act. This limits the applications of their work.

2.4.2.3 SmartMail

The SmartMail system (Corston-Oliver et al., 2004) is probably the most mature previous work on utterance-level request classification. SmartMail, shown in Figure 2.8, attempted to automatically extract and reformulate action items from email messages for the purpose of adding them to a user's to-do list.

The SmartMail system employed a series of deep linguistic features, including phrase structure and semantic features, along with word and part-of-speech n-gram features. The authors found that word n-grams were highly predictive for their classification task, and that there was little difference in performance when the more expensive deep linguistic features were added. This result is consistent with the results published by Cohen, Carvalho and Mitchell. Based on this insight, we do not employ the extra resources required to model deep linguistic features in our classification work in Chapters 4 and 5.

The results reported reveal only the aggregate performance across all classes, which involves a mix of both form-based classes (such as signature content address lines and URL lines), and intent-based classes (such as requests and promises). Results were also published as precision/recall curves, without explicitly identifying the results or peak F-score achieved. It is thus very difficult to directly compare the results with other systems. From the published graphs, however, it appears that peak F-score across all classes of speech acts, would be around 0.4 for sentence classification and 0.6 for message classification, which highlights the difficulties of automatic speech act classification at the utterance level.

The experiments with SmartMail were performed over a large corpus of messages that are not available for use by other researchers. In contrast, in our own work, we use messages from the widely-available Enron email corpus (Klimt and Yang, 2004) for our experiments.

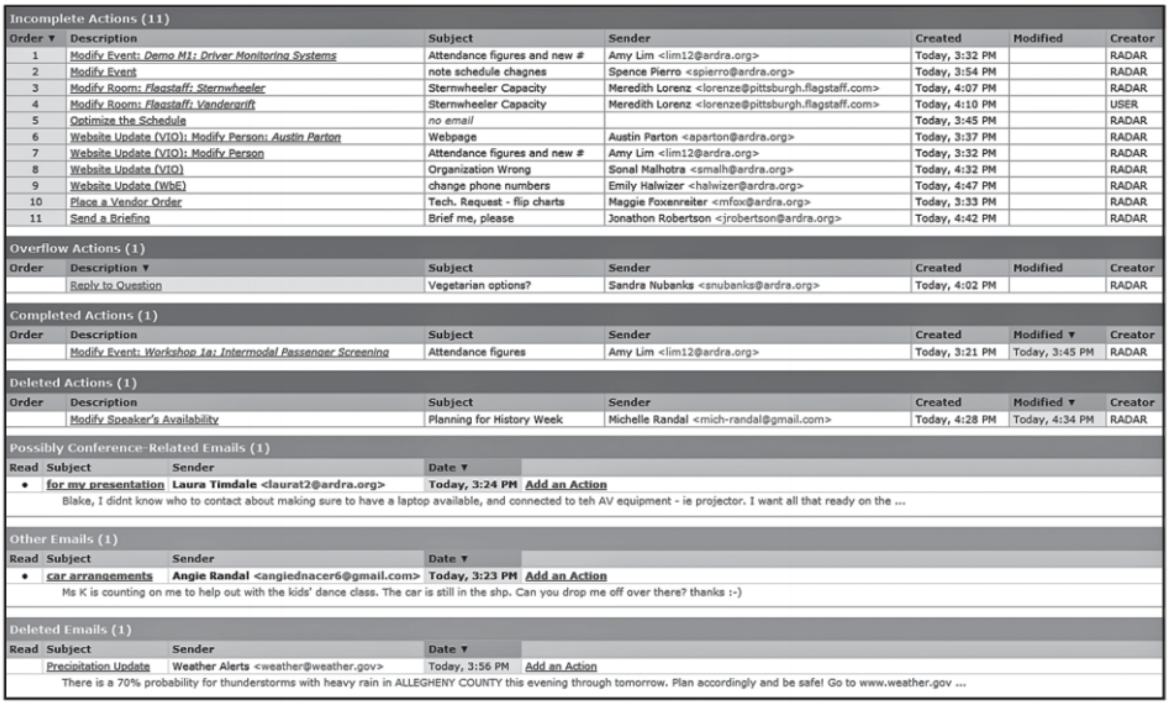
2.4.2.4 Semanta

The Semanta system is another prototype system that encompassed an ambitious approach to providing support for workflows in email (Scerri et al., 2009). Within that program of work, attempts were made to automate the identification of speech acts in email messages. The only published results use a rule-based approach which demonstrates relatively poor levels of accuracy. Specifically, the authors report an F-score of 0.58 in detecting requests across a small sample of 116 email messages (Scerri et al., 2010). While their plans for future work included employing statistical approaches for the classification task, no results from this work have yet been published.

One things that we take away from the Semanta work is a reinforcement of the difficulty of manually creating rules that capture the nuances of the wide variety of forms in which requests and commitments can be expressed. The low F-score scores achieved across the evaluations to date for requests will likely be lower for commitments, given the lack of a commissive mood, and thus their looser coupling to syntactic forms. We discuss this in more detail in Section 3.1.3.3. The relatively low F-scores in the results of this work support the approach of using relatively shallow features in a machine learning framework, rather than attempting to encode deep linguistic knowledge. As we discuss in Chapters 4 and 5, this is the approach we take for our computational experiments.

2.4.2.5 RADAR

A similarly ambitious program of work that touches on actionable email is the RADAR system, developed over many years (Freed et al., 2008). Part of the RADAR system is an email classifier that attempts to extract tasks from email messages, supported by a domain-specific ontology and other resources. One view of RADAR's Action List (Faulring et al., 2009) is shown in Figure 2.9, populated by the email classifier. It is immediately apparent that RADAR is designed to work within known domains of tasks, such as conference organisation, rather than being focused on open domain task recognition. Although most evaluation has focused on extrinsic evaluation of the



Incomplete Actions (11)

Order ▼	Description	Subject	Sender	Created	Modified	Creator
1	Modify Event: *Demo M1: Driver Monitoring Systems*	Attendance figures and new #	Amy Lim <lim12@ardra.org>	Today, 3:32 PM		RADAR
2	Modify Event	note schedule chagnes	Spence Pierro <spierro@ardra.org>	Today, 3:54 PM		RADAR
3	Modify Room: *Flagstaff: Sternwheeler*	Sternwheeler Capacity	Meredith Lorenz <lorenze@pittsburgh.flagstaff.com>	Today, 4:07 PM		RADAR
4	Modify Room: *Flagstaff: Vandergrift*	Sternwheeler Capacity	Meredith Lorenz <lorenze@pittsburgh.flagstaff.com>	Today, 4:10 PM		USER
5	Optimize the Schedule	*no email*		Today, 3:45 PM		RADAR
6	Website Update (VIO): Modify Person: *Austin Parton*	Webpage	Austin Parton <aparton@ardra.org>	Today, 3:37 PM		RADAR
7	Website Update (VIO): Modify Person	Attendance figures and new #	Amy Lim <lim12@ardra.org>	Today, 3:32 PM		RADAR
8	Website Update (VIO)	Organization Wrong	Sonal Malhotra <smalh@ardra.org>	Today, 4:32 PM		RADAR
9	Website Update (WbE)	change phone numbers	Emily Halwizer <halwizer@ardra.org>	Today, 4:47 PM		RADAR
10	Place a Vendor Order	Tech. Request - flip charts	Maggie Foxenreiter <mfox@ardra.org>	Today, 3:33 PM		RADAR
11	Send a Briefing	Brief me, please	Jonathon Robertson <jrobertson@ardra.org>	Today, 4:42 PM		RADAR

Overflow Actions (1)

Order	Description ▼	Subject	Sender	Created	Modified	Creator
	Reply to Question	Vegetarian options?	Sandra Nubanks <snubanks@ardra.org>	Today, 4:02 PM		RADAR

Completed Actions (1)

Order	Description	Subject	Sender	Created	Modified ▼	Creator
	Modify Event: *Workshop 1a: Intermodal Passenger Screening*	Attendance figures	Amy Lim <lim12@ardra.org>	Today, 3:21 PM	Today, 3:45 PM	RADAR

Deleted Actions (1)

Order	Description	Subject	Sender	Created	Modified ▼	Creator
	Modify Speaker's Availability	Planning for History Week	Michelle Randal <mich-randal@gmail.com>	Today, 4:28 PM	Today, 4:34 PM	RADAR

Possibly Conference-Related Emails (1)

Read	Subject	Sender	Date ▼	
•	**for my presentation**	**Laura Timdale** <laurat2@ardra.org>	**Today, 3:24 PM**	**Add an Action**

Blake, I didnt know who to contact about making sure to have a laptop available, and connected to teh AV equipment - ie projector. I want all that ready on the ...

Other Emails (1)

Read	Subject	Sender	Date ▼	
•	**car arrangements**	**Angie Randal** <angiednacer6@gmail.com>	**Today, 3:23 PM**	**Add an Action**

Ms K is counting on me to help out with the kids' dance class. The car is still in the shp. Can you drop me off over there? thanks :-)

Deleted Emails (1)

Read	Subject	Sender	Date ▼	
	Precipitation Update	Weather Alerts <weather@weather.gov>	Today, 3:56 PM	Add an Action

There is a 70% probability for thunderstorms with heavy rain in ALLEGHENY COUNTY this evening through tomorrow. Plan accordingly and be safe! Go to www.weather.gov ...

FIGURE 2.9: The RADAR System showing the Action List.

RADAR system as a whole, results from one published study demonstrate a macro-averaged F-score of 0.67 for the email classification component (Freed et al., 2008), which uses a regularised logistic regression suite of classifiers for this task. The system attempts to identify eight task types.

Because we aim to work on open domain email, the nature of our approach is naturally different to that of the RADAR program. We do not, for example, attempt to model the semantic meaning of utterances in terms of formal ontologies. We do, however, draw encouragement from the fact that RADAR has been extensively evaluated (Steinfeld et al., 2007; Faulring et al., 2010) and that the results have demonstrated that RADAR can provide useful support for assisted task management in email, the same style of support that we seek to offer in a more open domain.

2.4.2.6 Email Sentence Classification

Khoo, Marom, and Albrecht (2006) focus their work on sentence-level speech act detection in a corpus of real-world help desk email messages. They explore a range of classification algorithms, feature selection and pre-processing techniques for classifying sentences using a subset of the DAMSL Switchboard taxonomy.

They found that Support Vector Machine (SVM) classification generally outperformed the other algorithms that they experimented with, and that SVMs were largely insensitive to feature selection. Importantly, they found that, as sentences contain less textual information than documents, in pruning the feature space one needs to be very careful not to eliminate strong discriminative features, especially when there is a large class distribution skew. They also observed that lemmatisation and stopword removal proved detrimental to classifier accuracy, in contrast to the demonstrated useful dimensionality reduction they provide in document-level text classification. These are insights that we leverage in our own work in this thesis.

Ulrich et al. (2009) have also experimented with speech act classification in email, for the purposes of email summarisation. They found that the automatically generated labels they used at the message level were too coarse grained to be useful for summarisation, and that manually generated speech act labels at the utterance level were indeed useful for summarisation. This is encouraging for our own work, where we see one application of our sentence-level classifiers being the enabling of action-oriented summaries of email messages, building upon the early work done by the Pyam system, discussed in Section 2.4.2.1.

2.4.2.7 TOEIC Assessment

Recent work by De Felice and Deane (2012) focuses on building classifiers that attempt to automatically identify speech acts in second language learners' email messages, specifically in the context of a Test of English for International Communication (TOEIC) writing test for learners of English as a Foreign Language (EFL). This work is presented as the first step towards automating the scoring of these test items.

The corpus is derived from an exercise where students are instructed to write an email containing specific actions, such as asking for information or making a request—items that can easily be mapped onto expected speech act categories. A key criteria for scoring each student's answer is that all speech acts required by the prompt should be present in the answer text, making automatic speech act identification a useful component of automated scoring for this task.

De Felice and Deane use the Verbal Response Modes (VRM) taxonomy of speech acts discussed in Section 2.3.1 as their categories, and report accuracy up to 79.28% across six VRM categories. The six categories they employ are:

- Advisement (AA): Requiring action from the hearer, or a change in their mental state;
- Disclosure (DD): First person statements sharing thoughts that cannot be verified;
- Commitment/Factual Statement (DE): First person statements conveying verifiable facts;
- Indirect Advisement (QA): Polite or indirect requests or orders;
- Simple Question (QQ): Requests for information, not action; and
- Other Statements: Exclamations, third person statements, and any other utterances.

It is particularly interesting to note the conflated Commitment/Factual Statement category. De Felice and Deane have clearly identified the same weakness we saw with

VRM in identifying commitments, specifically in its inability to easily distinguish commitments from other types of utterances.

The authors readily acknowledge that their results are aided by the prevalence of direct speech acts, a consequence of the typically more limited repertoire of syntactic forms that EFL students are confident in using. This is confirmed by a significant drop in accuracy to 65.19% when their classifiers are run across email data from native English speakers.

As a general observation, De Felice and Deane note that requests, both in the form of imperatives and questions, are the most clearly identifiable. Conversely, the lower precision scores for commitments reinforce the more difficult nature of automatically recognising these acts.

2.4.2.8 Unsupervised Classification

Another recent thread of emerging research is the use of unsupervised and semi-supervised learning techniques for speech act recognition in textual conversations.

Jeong, Lin, and Lee (2009) developed a semi-supervised method for automatic speech act recognition in email and forum conversations. To overcome the lack of labelled data in these two genres, they apply domain adaptation techniques using labelled data from the DAMSL Switchboard and the Meeting Recorder Dialogue Act projects.

Ritter, Cherry, and Dolan (2010) were among the first to apply unsupervised methods to speech act recognition in textual conversation, using data from Twitter. They made some progress towards identifying the sequence patterns of tweets across conversations, though did not tackle the challenge of applying appropriate category labels to each message. Joty, Carenini, and Lin (2011) build on this work and apply similar unsupervised techniques to recognising speech acts in forums and email conversations. They find that the sequence dependencies can be more effectively learned by taking the conversational structure into account than by relying on structural or lexical similarity.

These techniques are still developing, and it is not yet clear how such a context-sensitive activity like learning the pragmatic meaning of utterance can be approached in an unguided, fully unsupervised manner. Given the ever-increasing volumes of textual conversation and the high cost of creating gold-standard, manually annotated corpora of such phenomena, however, it is likely that unsupervised and semi-supervised techniques for speech act recognition will receive increasing attention in future research.

2.4.3 A Summary of Systems for Task Management in Email

Collectively the systems surveyed above provide a range of insights that inform our own approach to building systems that assist with email task management. From the reviews of the Coordinator and TaskMaster systems, we recognise the need to minimise the manual effort required from users to identify tasks to encourage user adoption. We approach this goal by integrating automatic classifiers to alleviate the burden of manual obligation act identification. We describe these classifiers in Chapters 4 and 5, and describe how they are integrated into user-facing email software in Chapter 6.

Additionally, we recognise that it is important for users to maintain control over which requests and commitments are identified in their email, as not all message types and obligations are of equal relevance to users. As we describe in Chapter 6, we allow users to vet and confirm or reject automatic classifications, as well as to manually modify, create or delete obligation acts, as appropriate. As evaluation of the Coordinator system concluded, there are times that people do not want to remove all ambiguity from their messages by making their intentions completely explicit. We aim for our system to be flexible enough to provide useful support, but also allow the user to control when and where that support is deployed.

From the automated systems reviewed in this section, we recognise the need to provide support to users, regardless of the specific topics or workflows that emerge in their mailbox. Instead of encoding support for specific workflows, we aim to provide maximum flexibility by identifying requests and commitments in content from any domain. We start with a set of generic classifiers and tune these over time to suit the particular language styles and topics in each user's inbox, based on the user's implicit and explicit feedback. This avoids the need to manually encode rules, as in (Khosravi and Wilks, 1999; Scerri et al., 2010), which struggle to identify the wide variety of forms in which requests and commitments can be expressed across an open set of possible domains. Related to this, we acknowledge that the additional complexity of using deep-linguistic information does not necessarily improve the accuracy of speech act recognition (Corston-Oliver et al., 2004; Scerri et al., 2010). Instead, our approach is to be use mostly syntactic and lexical information for classification, features have been used with wide success for related problems.

2.5 Summary

Email is a critical medium through which much task-based communication is executed in the workplace. Previous research into the nature of such communication informs us that requests and commitments are the fundamental building blocks by which such 'conversations for action' take place. It is for this reason that, for the remainder of this thesis, we focus our attention on the nature of these two classes of speech acts, and on building systems to automatically identify these acts in email messages.

As we have highlighted through our review of related literature in this chapter, attempts to categorise and identify speech acts in computer-mediated communication such as email are not new. Unfortunately, despite the significant body of work that we have reviewed, none of the existing taxonomies and definitions are completely suitable for our needs, leading us to use a combination of empirical and theoretical techniques to derive our own definitions of requests and commitments in Chapter 3.

Similarly, we have seen that the idea of providing better support for task-related conversations in email is also not new. Although commercial email systems provide poor support for task management, there are many prototype systems that have tried to innovate in this space. Many of these systems have done so without formalising or deeply understanding the complexity of the requests and commitments that they seek to identify and surface. Again, this is a key limitation that we address in Chapter 3.

Additionally, we take important lessons from the collection of these systems to inform the design and implementation of our own approaches and systems for identifying task-related content in workplace email.

"I live in the inbox and that is kind of my to-do list. I'll keep things in there ... that I have to deal with"

Email user, quoted in (Whittaker and Sidner, 1996, p. 281)

3 An Examination of Requests and Commitments in Workplace Email

As discussed in Chapter 2, studies have repeatedly highlighted that communicating actionable information and delegating tasks are important and frequent uses for email. Our research focuses on such task-oriented email usage, and specifically on identifying REQUESTS (directive speech acts) and COMMITMENTS (commissive speech acts). As we have begun to see from our discussion of Speech Act Theory in Chapter 2, the expression and detection of such phenomena are frequently complex. The key contribution of this chapter is an examination and account of how requests and commitments acts are conveyed in real-world, workplace email communication. We focus on the way these acts function and the manner in which they are realised and interpreted. This builds the foundation for our work on automating the recognition of requests and commitments in email communication that is presented in Chapters 4 and 5.

We begin this chapter by carefully scoping in Section 3.1 how we identify utterances which convey some obligation for future action in the form of requests and commitments. In this way, we start to define our focus on identifying the utterances that are involved in exchanging, delegating and otherwise creating obligations for future action. At its core, the task of identifying requests and commitments is one of recognising an email author's intention. This makes our task an instance of the more general problem of recognising the speech acts (or the intentions) behind an author's utterances, a class of problems which are by their very nature subjective and context-sensitive. We work through this ambiguity to present definitions for identifying both requests and commitments in email text.

In Section 3.2 we support and refine our definitions with empirical evidence. We provide an overview of a series of annotation experiments we performed to observe and analyse how people make requests and commitments in real-world email messages.

The results of our annotation experiments clearly demonstrate the significant complexity and context-sensitivity that is involved in the way people make requests and commitments. This makes the reliable identification of requests and commitments a challenging task. To address this difficulty, in Section 3.3 we identify, categorise and analyse a range of phenomena, such as the use of phatic obligation acts, and acts that function simultaneously as requests and commitments, which have led to disagreement among our independent human annotators when asked to identify requests and commitments.

The theoretical and empirical observations and analysis in this chapter lay the foundation for the rest of this thesis, and for other researchers exploring task-based language in computer-mediated communication (CMC) in general, and email in particular. Having a firm grasp of what constitutes a request or a commitment in real-world email usage is obviously required in order to design and build systems for automatically identifying such actionable content. The analysis in this chapter provides such a grounding, facilitating our work on automating the identification of requests and commitments in later chapters.

3.1 The Nature of Requests and Commitments in Email

Consider the following utterance in an email from one workplace email user, Finn, to his colleague, Ronan:

(3.1) Finn: *Please let me know if you plan to attend the meeting.*

Finn's utterance acts as a REQUEST for a response from Ronan. Now, consider Ronan's response:

(3.2) Ronan: *Yes, I will be there.*

In this case, Ronan's response acts as a COMMITMENT, placing an obligation for future action on Ronan to turn up to the meeting.

Our focus is on exploring how such requests and commitments are used in workplace email. We are not focused on the propositional content of what is being requested or promised; instead we focus on the common abstract intentions that lie behind all such utterances, regardless of the topic of conversation. Speech Act Theory, as we discussed in Section 2.2.1, provides just such abstractions; Searle would call our requests directive speech acts, and our commitments commissive speech acts. Both these classes of speech acts are concerned with utterances that convey or impose obligations for future action on the author or recipient. An interesting property in email is that requests and commitments utterances tend to mirror each other, in terms of whom the obligation for future action is placed upon: requests place an obligation on a recipient of a message, and commitments commonly place an obligation on the sender.

To illustrate this symmetry in requests and commitments, consider a second request, this time sent in an email from Ronan to Finn:

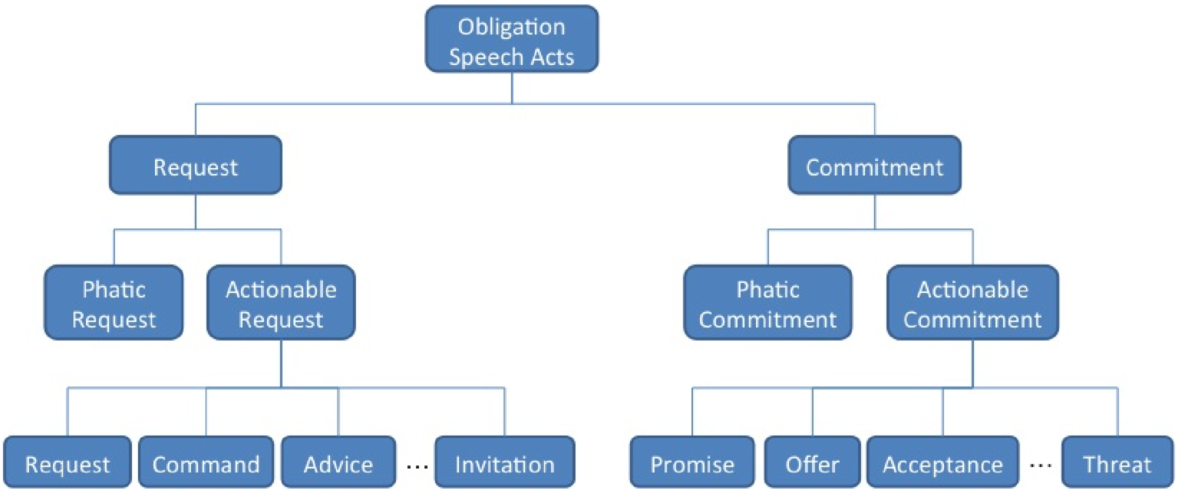


Figure 3.1: A taxonomy of request and commitment speech acts.

(3.3) *Finn, please send me the latest version of the document.*

This request places an obligation on Finn to send the latest version of the document, or otherwise respond to Ronan's request. From Ronan's perspective, there's an obligation pending on Finn to act on his request. That obligation is essentially the same as the obligation created by an incoming commitment sent by Finn, as in the following utterance:

(3.4) *I'll send the latest version to you today.*

The same mirroring is seen from Finn's perspective. The utterances in both Example (3.3) and Example (3.4) place an obligation on him to send the latest version of the document to Ronan.

This mirroring is important, since it explains why email users need to track obligations in both incoming and outgoing email messages: an incoming request received by Ronan can create an obligation for action for him in the same way as Ronan sending a commitment to do something. The same mirroring is seen for incoming commitments and outgoing requests: both are obligations that Ronan does not have to action himself, but in both cases Ronan might well want to follow up on their execution with the responsible party.

Figure 3.1 illustrates a taxonomy of requests and commitments in a more visual way. Both types of speech act are focused on future action. We use the term obligation acts as a superset of both requests and commitments to encapsulate all speech acts that create or confirm an obligation for someone to take action. As can be seen in this figure, our use of the term request is much broader than the single sub-category of directive speech acts sometimes referred to by linguists; we refer to all types of directive acts (e.g., commands, suggestions, invitations and requests[1]) as requests. Similarly, we refer to all commissive acts (e.g., offers, promises, acceptances, refusals and threats) as commitments.

We see the same set of sub-classes for both requests and commitments in Figure 3.1. The actionable sets of acts refer to the requests and commitments which actually place an obligation on either the sender or a recipient to take action. These include requests like *Please send me the document*, and commitments like *I'll send the document this afternoon.* The phatic acts refer to speech acts that resemble requests or commitments, either via their literal meaning or by convention, yet place very little or no obligation on anyone to act. Common phatic requests include utterances like *Let me know if you have any questions*; common phatic commitments include utterances like *Talk to you soon* that convey little, if any, commissive obligation on the sender.[2] We discuss phatic requests and phatic commitments further in Section 3.3.1.

3.1.1 Understanding Requests and Commitments

As discussed in Chapter 2, a body of previous research has defined a range of speech act taxonomies. Despite a detailed survey of this existing work, we have not found any single taxonomy directly applicable for our purposes, due to a range of different limitations (Section 2.3). We briefly revisit a few of the more pertinent issues below.

One common limitation is a lack of guidance around whether conditional requests and commitments should be identified as obligation acts. Conditional requests and commitments are those where the obligation to perform the associated action is contingent upon some other state holding true, as in Example (3.5) below.

(3.5) *If we finish early, can you please let Greta know?*

We explore conditional obligation acts in more detail in Section 3.1.7.

Another constraint is that many of the taxonomies are constructed based on the use of short, simple definitions or a sample of canonical examples to define each of their speech act categories. Leuski (2004), for example, defined his request action category with a single example utterance: *Please reserve a room.* We experimented with using a similarly simple approach in our early annotation experiments (Lampert, Paris, and Dale, 2007), and the resulting levels of disagreement between human annotators clearly demonstrated that such definitions lack the detail and clarity required for reliable and consistent identification and classification of the complex requests and commitments we find in real-world email. Other researchers have since recognised similar problems with relying on unrealistically simple definitions (Scerri et al., 2008). We discuss some of the specific issues that arose in our early annotation experiments in Section 3.2.3.

[1] Note that here we use the word 'request' in its narrower linguistic meaning, as a specific sub-type of directive speech act. Through the rest of the thesis, we use the word 'request' to mean the class of all directive speech acts.

[2] The phatic acts we refer to are very similar to Clark's (1997, p. 318) ostensible communicative acts.

To provide a set of guidelines for our manual and automated annotation tasks, in this chapter we carefully define how we identify speech acts that are (and those that are not) requests and commitments. For us, the ontological foundation of our request and commitment definitions is the notion of an ACTION. Actions are carried out by AGENTS, which are the people, organisations or other groups involved. Importantly, we consider the linguistic realisation of a request or commitment to be distinct from the request or commitment act itself. This allows us to adopt the traditional linguistic distinction between DIRECT SPEECH ACTS and INDIRECT SPEECH ACTS. A direct speech act is one where either the literal interpretation of the realisation matches the underlying speech act, or alternatively, where the realisation is a well-understood, conventional form of the underlying act. Examples of direct requests include:

(3.6) *Please send me the document.*

(3.7) *Details please.*

An indirect speech act is one where the literal interpretation of the surface form of the act does not match the underlying speech act. Examples tend to be context sensitive, but could include the following act as a request to finish editing the document by Friday.

(3.8) *I'd love to have the document finished by Friday.*

Separating the realisation from the underlying act also allows us to talk about multiple realisations of the same underlying request or commitment within an email message. Consider, for example, the message in Figure 3.2.[3] This messages contains two separate utterances, shown as Example (3.9) and Example (3.10), that repeat and reword the same request for Bruce, namely for him to contact Sean.

(3.9) *I was just wondering where we are at now.*

(3.10) *If you get a chance, either e-mail me or give me a call.*

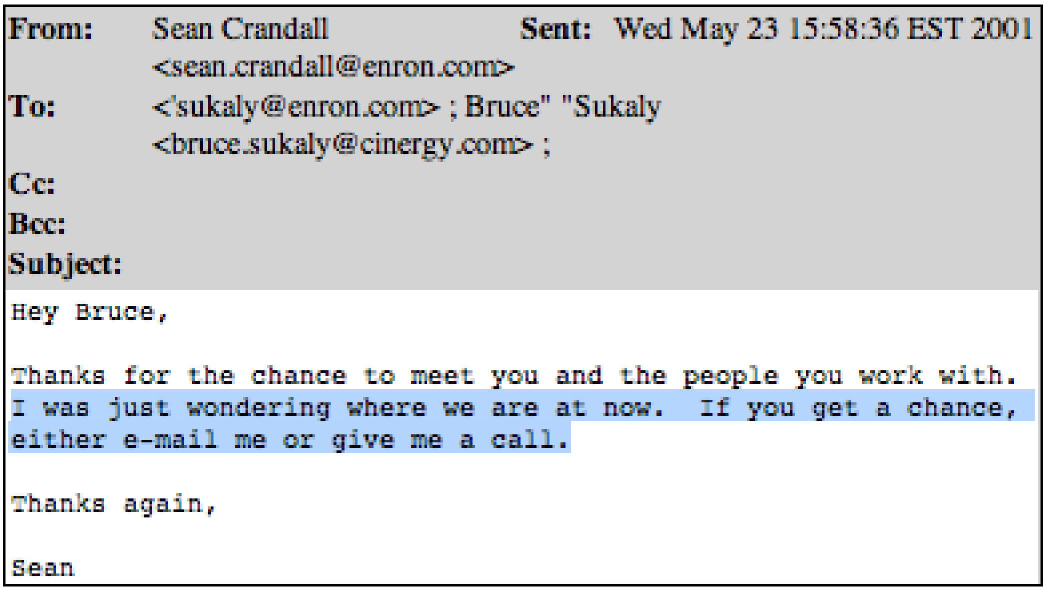
We elaborate on these definitions below, focusing on requests in Section 3.1.2 and commitments in Section 3.1.3. Our definitions are then supplemented by deep analysis of particular cases of ambiguity and complexity that we explore in detail in Section 3.3.

3.1.2 Defining Requests in Email

We define any utterance from an email sender that places an obligation for future action on one or more email recipients as a request. The associated action may be to:

1. Perform a physical action—e.g.,

[3] Example email messages used are drawn from the Enron email dataset. We use the data in its publicly available, unanonymised form, noting that particularly sensitive messages were redacted before the corpus was publicly released.



From: Sean Crandall <sean.crandall@enron.com> **Sent:** Wed May 23 15:58:36 EST 2001
To: <'sukaly@enron.com> ; Bruce" "Sukaly <bruce.sukaly@cinergy.com> ;
Cc:
Bcc:
Subject:

```
Hey Bruce,

Thanks for the chance to meet you and the people you work with.
I was just wondering where we are at now.  If you get a chance,
either e-mail me or give me a call.

Thanks again,

Sean
```

Figure 3.2: An email that contains two distinct surface realisations of the same underlying request, as highlighted.

(3.11) *Please replace the toner in the downstairs printer.*

2. Perform some speech act—e.g.,

(3.12) *Please let me know if this is ok with you.*

Where a response by means of some speech act is required, the response is frequently required to be communicated back to the requestor, but may instead need to be directed to other identified agents. An example of such a request is:

(3.13) *Please let Michelle know if you're coming.*

As well as the distinction between physical action and communicative action, we can also distinguish between immediate and future actions. Requests for future actions may involve the recipient scheduling, rather than actually performing an action. Consider, for example, the utterance in Example (3.14) that does not require any immediate action from the recipient, unlike the utterance in Example (3.3).

(3.14) *Please send me the document on Friday.*

We discuss this distinction further in Section 3.3.2.

In linguistic terms, requests are directive speech acts. The class of directive acts includes a range of specific speech acts such as:

(3.15) Requests[4]—e.g., *Can you please send me the document?*

(3.16) Commands – e.g., *Send me the document by 4pm.*

(3.17) Suggestions – e.g., *Why don't you send me the document?*

(3.18) Questions—e.g., *Can I get a copy of the document?*

(3.19) Advice—e.g., *I think you should just send the current version to Robert.*

We consider each of these classes of directive acts to be requests. Labov and Fanshel (1977) focused specifically on analysing requests in institutional discourse, and found a wide range of request types in use, including:

(3.20) Requests for action—e.g., *Please lock the door at 5pm.*

(3.21) Requests for information—e.g., *How much capacity do we have remaining?*

(3.22) Requests for permission—e.g., *May I finish early to attend the training?*

(3.23) Requests for confirmation—e.g., *You're coming to the meeting, right?*

(3.24) Requests for agreement—e.g., *Agreed?*

(3.25) Requests for evaluation—e.g., *Let me know what you think of him.*

(3.26) Requests for interpretation—e.g., *What are your thoughts on the issue?*

Again, we include all of these within the scope of requests that we seek to identify in workplace email. Some linguists (for example, Sinclair and Coulthard (1975)) have distinguished between speech acts that require a physical response from those that require a verbal or information response. However, we follow Searle's (1969) original approach and consider utterances requiring either physical or speech act responses to both be classes of requests. We thus explicitly include questions seeking an informational response as requests, since they represent an attempt by the sender to elicit an action from the recipient, in the form of a speech act.[5] Given our focus on unconstrained, real-world email, we also consider the set of possible actions to be open, and thus do not attempt to enumerate the set of valid actions.

Also important to note is that the type of speech act required in response to a request is open; frequent acts that might be made in response to a request include, but are not limited to:

(3.27) Informing—e.g., *I think it costs about $10.*

[4] As noted earlier, *request* in this context refers to a specific sub-type of directive act. Throughout the rest of this thesis, we use the term request to mean the set of all types of directive acts.

[5] Note, however, that not all questions function as requests, rhetorical questions being the most obvious example of non-request questions. Some researchers have claimed that email text exhibits more frequent use of rhetorical questions than other media (Crystal, 2001, p. 122).

(3.28) Permitting—e.g., *Yes, you may leave early today.*

(3.29) Committing—e.g., *Sure, I'll see you at 10am tomorrow.*

(3.30) Accepting—e.g., *Thanks, I'd love some.*

(3.31) Refusing—e.g., *I don't need any, thanks.*

Our definition of a request is deliberately broad. We are interested in identifying all utterances in an email message that place obligation for current or future action on a user who receives that message.

3.1.2.1 The Representation of Requests

In terms of our ontology, we define a request as an utterance that places an OBLIGATION from one agent (the REQUESTOR) on another agent (the REQUESTEE) to carry out the requested action. Formally, a request (R) is represented as:

R = ⟨Action, Requestor, Requestee⟩

Here's how we represent the utterance in Example (3.11) (*Please replace the toner in the downstairs printer.*) when sent in an email from Ronan to Finn:

R = ⟨Action=Replace Printer Toner, Requestor=Ronan,
Requestee=Finn⟩

As well as representing the underlying request act, we also separately represent each realisation of a request act. We call each separate realisation an UTTERED REQUEST (UR), represented as:

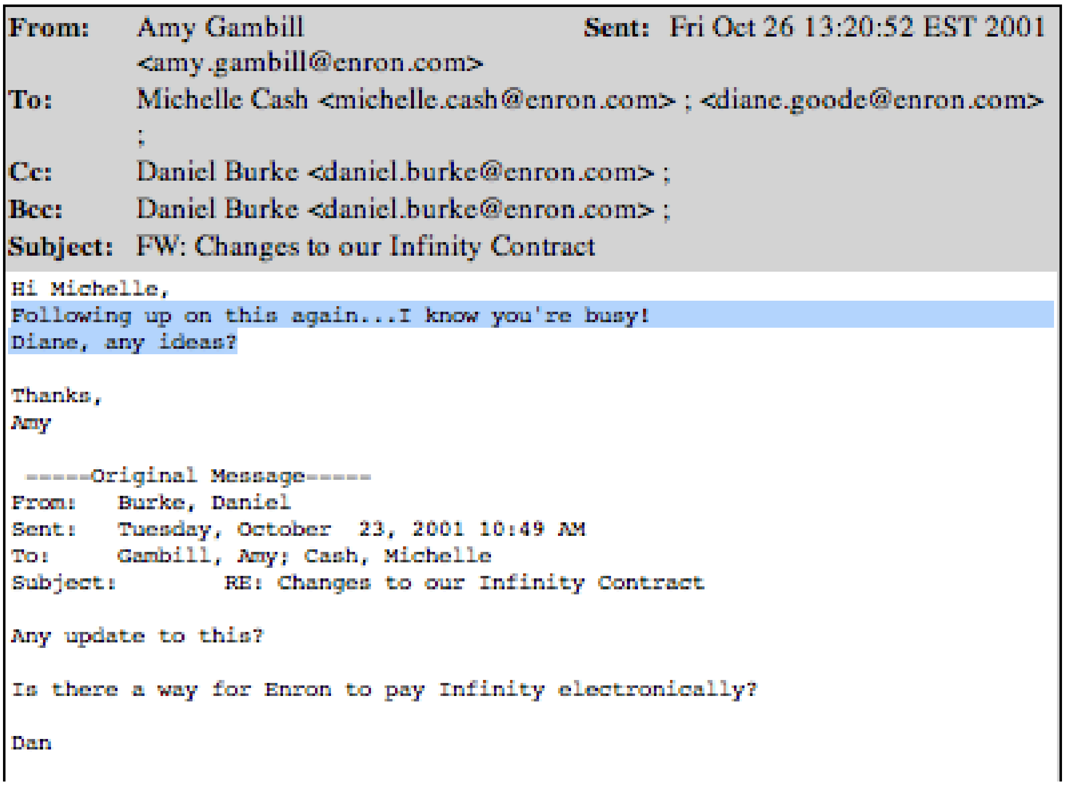
UR = ⟨Direct/Indirect, Utterance, Request⟩[6]

As we saw in the message in Figure 3.2, senders sometimes repeat and reword the same request within the same message. The two uttered requests from the message in Figure 3.2 would be represented as shown below in Example (3.32). Note that both uttered requests refer back to the same underlying request.

(3.32) R1 = ⟨Action=Phone/Email, Requestor=Sean Crandell,
Requestee=Bruce Sukaly⟩
UR1 = ⟨Indirect, *I was just wondering where we are at now*, R1⟩
UR2 = ⟨Direct, *if you get a chance, either e-mail me or give me a call*,
R1⟩

A different case is illustrated in the email message in Figure 3.3, which also contains two realisations of a similar underlying request. In this case, however, because the two uttered requests are addressed to different recipients, they are, by definition, distinct. The first uttered request is modelled in Example (3.33).

[6] We use the term 'utterance' as a size-independent label for a text fragment used to convey a request or commitment.



From: Amy Gambill <amy.gambill@enron.com> **Sent:** Fri Oct 26 13:20:52 EST 2001
To: Michelle Cash <michelle.cash@enron.com> ; <diane.goode@enron.com> ;
Cc: Daniel Burke <daniel.burke@enron.com> ;
Bcc: Daniel Burke <daniel.burke@enron.com> ;
Subject: FW: Changes to our Infinity Contract

```
Hi Michelle,
Following up on this again...I know you're busy!
Diane, any ideas?

Thanks,
Amy

 -----Original Message-----
From:    Burke, Daniel
Sent:    Tuesday, October  23, 2001 10:49 AM
To:      Gambill, Amy; Cash, Michelle
Subject:         RE: Changes to our Infinity Contract

Any update to this?

Is there a way for Enron to pay Infinity electronically?

Dan
```

Figure 3.3: An email containing two realisations of the same underlying request, one declarative and one interrogative, addressed to different recipients.

(3.33) R = ⟨Respond, Amy Gambill, Michelle Cash⟩
UR = ⟨Indirect, *following up on this again … I know you're busy!*, R⟩

The request in Example (3.33) is addressed explicitly to Michelle, and while phrased as a declarative act, is clearly intended to function as a request. Because the literal meaning does not belie its function as a request, we consider this realisation to be an indirect speech act. Our own analysis of workplace email confirms previous studies that have highlighted more frequent use of indirect speech acts in email than in other media (Hassell and Christensen, 1996).

The second request in Figure 3.3 is represented in Example (3.34). This request realisation, phrased as an interrogative act, also clearly functions as a request for the addressee, in this case, Diane, one of the two primary recipients of the message.

(3.34) R = ⟨Respond, Amy Gambill, Diane Goode⟩
UR = ⟨Direct, *Diane, any ideas?*, R⟩

The two required actions represent two distinct instances of the same type of response, to be completed by two different recipients. The requests in Examples (3.33) and (3.34) also demonstrate how different utterances in the same message are often addressed to different audiences, an issue we return to in Section 3.1.4.

3.1.2.2 Common Request Surface Forms

There are three surface forms that are most commonly used to express requests:

1. Imperative
2. Interrogative
3. Declarative

Of course, in addition to these, there are innumerable forms in which an utterance may represent an indirect request in some context. In this section, we do not attempt to produce an exhaustive set of request forms (as noted earlier, this would be impossible due to the loose coupling between form and function), but rather we aim to describe and exemplify the major forms used for issuing requests in email text.

Imperative Requests

The IMPERATIVE form conveys a sender's intention to influence a recipient's behaviour or actions. Commonly, the imperative form is used to express direct commands or requests. It can, however, also be used for other speech acts, such as wishing (e.g., *Have a great weekend!*).

For requests, there are at least four common classes of imperative usage:

1. **Base Form Verb** — where a bare, imperative verb with no subject is used. e.g.,

   (3.35) *Send it!*

2. **You + Imperative** — where the imperative command is prefixed by *you.* e.g.,

   (3.36) *You send it.*

3. **Let's + Imperative** — where the imperative command is prefixed by *Let's.* e.g.,

   (3.37) *Let's send it.*

4. **Embedded *if* clause** — where an *if* clause containing the request is embedded within a sentence. This form may also occur with the word *whether.* e.g.,

   (3.38) *See if you can send it.*

   (3.39) *Let's just see if you can send it.*

   (3.40) *Let me know whether you can send it.*

Interrogative Requests

Questions are the most prototypical type of request expressed using the INTERROGATIVE form. There are two common subtypes of interrogative usage:

1. **Modal Interrogative** — where the question is phrased with a modal verb. e.g.,

(3.41) *Can you send it?*

2. **Non-modal Interrogative** — where the question is phrased using other question words e.g.,

   (3.42) *What about if you just send it?*

Declarative Requests

Although not a canonical form, we also frequently see requests conveyed as DECLARATIVES, as in Example (3.43) and Example (3.44).

(3.43) *Another thing I'd like to know is who else at Enron is listed as an officer of EPMI.*

(3.44) *We will require you to fill out a form.*

There various imperative and interrogative surface forms are highly indicative of requests. Declarative requests present an additional challenge when working at the surface level, since this form of utterance can also represent many other non-directive speech acts, such as simple statements about the world (e.g., *Popcorn is yellow*). The purpose of surveying these common request forms, and particularly in identifying additional patterns within the collection of imperative and interrogative forms, is to inform the statistical machine learning classifiers that we develop in Chapters 4 and 5. We encode information about many of these surface forms in the feature sets for those classifiers, based on the cues they provide for recognising requests using both lexical and syntactic patterns.

3.1.3 Defining Commitments in Email

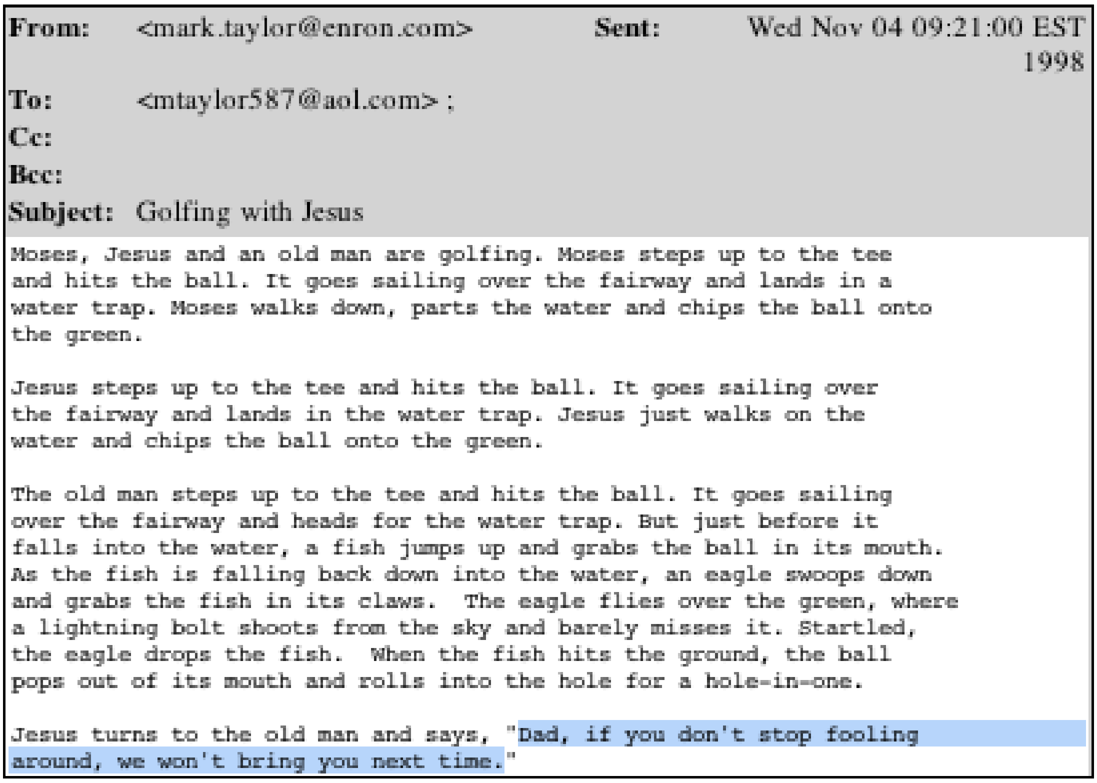
Consider the following utterance sent in an email from Greta to her colleague Archie:

(3.45) I will send you the latest version of the document by 5pm today.

This utterance places an obligation on Greta to follow through and complete her promised future action of sending the latest version of the document by the specified deadline. This utterance is an example of a commitment.

We define a commitment as any utterance that places an obligation for future action on the email sender or other party. In linguistic terms, such acts are considered to be commissive speech acts. Searle succinctly defines commitments as "those illocutionary acts whose point is to commit the speaker (again in varying degrees) to some future course of action" (Searle, 1976, p11). Austin notes similarly that "the whole point of a commissive is to commit the speaker to a certain course of action" (Austin, 1962, p156). One point of variation is that we expand the scope of commitments beyond that defined by Austin and Searle to include utterances that commit other parties, as we discuss further in Section 3.1.3.2.

Example commissive acts include:



From: <mark.taylor@enron.com> **Sent:** Wed Nov 04 09:21:00 EST 1998
To: <mtaylor587@aol.com> ;
Cc:
Bcc:
Subject: Golfing with Jesus

Moses, Jesus and an old man are golfing. Moses steps up to the tee and hits the ball. It goes sailing over the fairway and lands in a water trap. Moses walks down, parts the water and chips the ball onto the green.

Jesus steps up to the tee and hits the ball. It goes sailing over the fairway and lands in the water trap. Jesus just walks on the water and chips the ball onto the green.

The old man steps up to the tee and hits the ball. It goes sailing over the fairway and heads for the water trap. But just before it falls into the water, a fish jumps up and grabs the ball in its mouth. As the fish is falling back down into the water, an eagle swoops down and grabs the fish in its claws. The eagle flies over the green, where a lightning bolt shoots from the sky and barely misses it. Startled, the eagle drops the fish. When the fish hits the ground, the ball pops out of its mouth and rolls into the hole for a hole-in-one.

Jesus turns to the old man and says, "Dad, if you don't stop fooling around, we won't bring you next time."

Figure 3.4: An email with a reported threat being used to convey humour.

(3.46) promises—e.g., *I'll send you the document today.*

(3.47) offers—e.g., *Would you like me to send you the document today?*

(3.48) threats—e.g., *If you don't help, I'll send the document to your boss!*

In focusing on workplace email communication, we include promises and offers as commitments. Threats are another interesting case, though quite rare in the workplace email we have analysed. The threats that we do see in our corpus typically do not convey a real obligation that is relevant for workplace task management, and are instead used as instruments of humour or irony, often in the form of reported speech. An example of such usage is shown in Figure 3.4. Threats of this nature function much more as phatic commitments than true commitments. We discuss the distinction between threats and promises in Section 3.3.3.3.

3.1.3.1 The Representation of Commitments

In terms of our ontology, we define a commitment as an utterance that places an obligation from one agent (the committor) on another agent (the committee) to perform some future action. In email communication, the committee is frequently, but not always, the sender. Formally, a commitment (C) is represented as follows:

$$\mathrm{C} = \langle \textsc{Action}, \textsc{Committor}, \textsc{Committee} \rangle$$

As with requests, we separately represent each realisation of a commitment act as an UTTERED COMMITMENT. Each uttered commitment (UC) is realised as either a direct or an indirect speech act, represented as:

$$\mathrm{UC} = \langle \textsc{Direct/Indirect}, \textsc{Utterance}, \textsc{Commitment} \rangle$$

Consider the commitment utterance in Example (3.46) in an email sent from Ronan to his colleague Finn. In that context, the utterance commits Ronan to sending the document, and is represented as follows:

(3.49) $\mathrm{C} = \langle \textsc{Action}{=}\textsc{Send Document}, \textsc{Committor}{=}\textsc{Ronan}, \textsc{Committee}{=}\textsc{Ronan} \rangle$
$\mathrm{UC} = \langle \textsc{Direct}, \textit{I'll send you the document today}, \mathrm{C} \rangle$

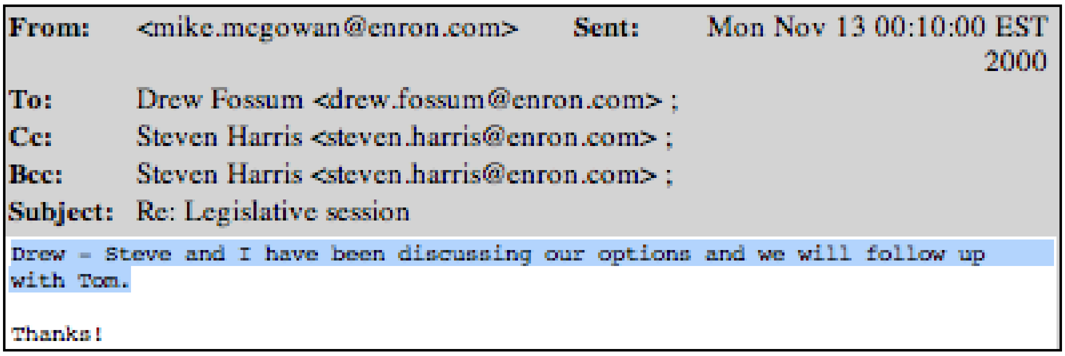
A useful, though not conclusive, diagnostic test for identifying a commitment is whether you might expect to find the future action as an entry in the responsible agent's task list or calendar. If it is unclear whose task list should contain the future action, then the committee for future action is not readily identifiable, and the we do not consider the utterance in question to be a commitment. There is little value in knowing that *someone* was going to perform a task if we do not know *whom* to follow up with. Consider the example below, as it occurs in an email from Ronan to Finn:

(3.50) *Someone will let you know.*

While this utterance is made by Ronan, he is not necessarily the person who will undertake the associated action. The distinction between unattributable commitments, such as this one, and actual commitments is, however, often not clear cut. In Example (3.50), Finn may still hold Ronan accountable as the committor, even if Ronan did not promise to take the action himself. Alternatively, consider the following utterance, which is also extracted from an email sent by Ronan to Finn.

(3.51) *Lunch will be served*

The utterance in Example (3.51) may or may not function as a commitment. Out of context, it is not clear that Ronan, or any other identifiable agent, is committed to serve or arrange the serving of lunch. If no agent can be identified for the associated action, then we would consider this to be an 'unassigned' action, and thus not a commitment. Ultimately, however, the wider context of the message might create a clear obligation for Ronan as the sender to follow through on the commitment, in which case the utterance would be considered a commitment, with Ronan identified as the committee and committor.



From: <mike.mcgowan@enron.com> Sent: Mon Nov 13 00:10:00 EST 2000
To: Drew Fossum <drew.fossum@enron.com> ;
Cc: Steven Harris <steven.harris@enron.com> ;
Bcc: Steven Harris <steven.harris@enron.com> ;
Subject: Re: Legislative session

```
Drew - Steve and I have been discussing our options and we will follow up
with Tom.

Thanks!
```

Figure 3.5: An email with an utterance that functions as both a third-party commitment (for Steve and Mike), and a request (for Steve, as a recipient of the message).

3.1.3.2 Commitments for Non-senders

In contrast to most other work that has considered commitments, including Austin (1962), Searle (1976) and more recent, email-specific work by Corston-Oliver et al. (2004), we consider utterances that place an obligation on a person, group of people, or organisation *other than the sender* to also be commitments. This definition represents an expansion of Austin and Searle's definitions of commitments, which were limited to commitments placed only upon the speaker/sender. An example of an utterance where the committee is someone other than the sender is shown in Example (3.52), taken from an email message sent from Ronan.

(3.52) *Michelle will send through the revised document by Friday*

This utterance would be represented as:

(3.53) C = ⟨Action=Send Document, Committor=Ronan,
Committee=Michelle⟩
UC = ⟨Direct, *Michelle will send through the revised document by Friday*, C⟩

Observation of the real-world Enron email corpus that we work with shows that commitment utterances such as these that place obligation on a committee other than the sender are relatively common. Such utterances, which we refer to as third-party commitments, can be quite nuanced in their intention. They can, for example, simultaneously function as a commitment for some recipients and a request for other recipients. An example of an utterance which demonstrates this duality is shown in Figure 3.5.

From the perspective of both Drew Fossum and Mike McGown, the utterance highlighted in Figure 3.5 and extracted as Example (3.54) is unambiguously a commitment. From the perspective of Steven Harris, however, the same utterance can function as a request, as represented in Example (3.55).

(3.54) C = ⟨Follow up with Tom, Committor=Mike McGowan, Committee=Mike McGowan, Steven Harris⟩
UC = ⟨Direct, *Drew — Steve and I have been discussing our options and we will follow up with Tom*, C⟩

(3.55) R = ⟨Follow up with Tom, Requestor=Mike McGowan, Requestee=Steven Harris⟩
UR = ⟨Indirect, *Drew — Steve and I have been discussing our options and we will follow up with Tom*, R⟩

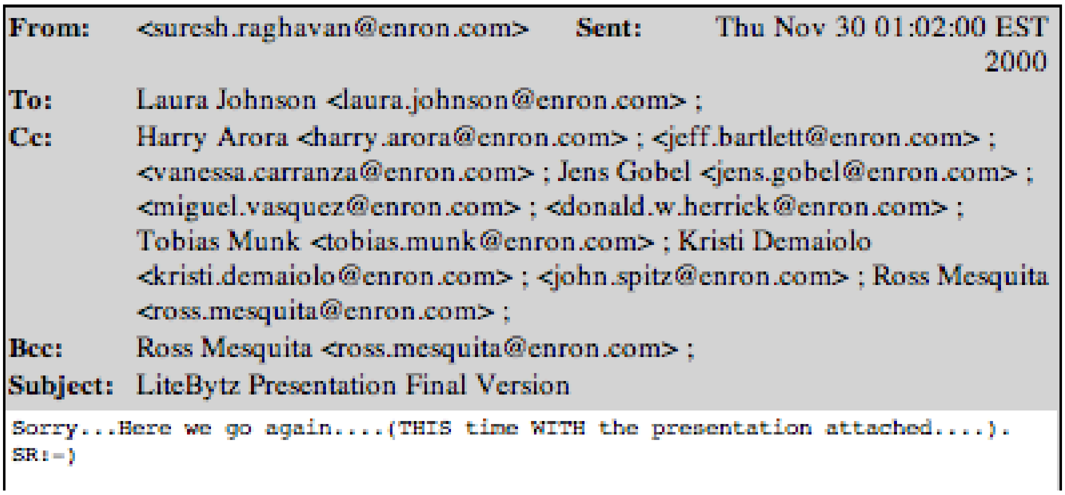
The request interpretation in Example (3.55) represents the utterance from the perspective of Steven Harris, where Mike McGown's emailed commitment to Drew Fossum (*... we will follow up with Tom*) can function as an indirect request asking or reminding Steve to follow up with Tom.

3.1.3.3 Common Commitment Surface Forms

Unlike requests, commitments are not realised syntactically in a particular sentence type or modality. One theory for this lack of a commissive mood, advanced in (Vanderveken, 1990), is that commitments are simply less important in communication than, say, requests. Additionally, Vanderveken notes that "speakers usually commit themselves indirectly by way of, for example, asserting literally that they will do something or that they intend to do something. In such cases, whenever they do not keep their commitment, they can always pretend they were only making a prediction" (Vanderveken, 1990, p. 109). The resulting lack of a prototypical syntactic form makes recognising commitments more difficult than requests. Even without a commissive form, however, mail senders can choose to make their commitment explicit by using a performative utterance with a commissive verb, e.g., *I promise to send you the document*, or *I agree to send you the document*. This combination of indirect and performative ways of uttering a commitment are considered sufficient for the purposes of communication.

One of the most common ways that email senders commit themselves to action is simply by writing a variant of the utterance *I will do it*. Even such an utterance can, however, be ambiguous; the sender may be making a commitment, but may also be simply expressing an intention (meaning *I will try to do it*), or making an unbinding prediction (meaning *I predict that I will do it*). This makes distinguishing a commitment from a prediction or a statement of intention challenging. Usually the context will make clear what the sender intended, but if there is uncertainty, the sender can resolve the ambiguity by using a performative verb (e.g., *I promise*) to mark their intention explicitly. In the email communication we have analysed, however, the use of such performatives is quite rare.[7] In order for *I will* to count as a commitment, it must make sense for the recipient to offer a challenge, such as *Do you promise?* with an appropriate response

[7] Across the 3 million sentences in the Enron corpus, there are less than 350 instances of the phrase *I promise* and less than 50 instances of the phrase *I predict*.



From:	<suresh.raghavan@enron.com>	**Sent:**	Thu Nov 30 01:02:00 EST 2000
To:	Laura Johnson <laura.johnson@enron.com> ;		
Cc:	Harry Arora <harry.arora@enron.com> ; <jeff.bartlett@enron.com> ; <vanessa.carranza@enron.com> ; Jens Gobel <jens.gobel@enron.com> ; <miguel.vasquez@enron.com> ; <donald.w.herrick@enron.com> ; Tobias Munk <tobias.munk@enron.com> ; Kristi Demaiolo <kristi.demaiolo@enron.com> ; <john.spitz@enron.com> ; Ross Mesquita <ross.mesquita@enron.com> ;		
Bcc:	Ross Mesquita <ross.mesquita@enron.com> ;		
Subject:	LiteBytz Presentation Final Version		

```
Sorry...Here we go again....(THIS time WITH the presentation attached....).
SR:-)
```

Figure 3.6: An example email message showing the three classes of email recipients: 'To', 'Cc', and 'Bcc'.

3.1.4 Considering the Email Audience

Email is a complex and flexible medium that can be used for both one-to-one and one-to-many communication. Additionally, the audience for a message can include both public and private participants in the conversation.

As a starting point, the audience for an email message can be identified from the header information in an email message. This information captures the set of recipients that the email sender explicitly identifies as the intended audience for their message. There are, however, some types of recipient addresses, such as mailing lists, that can obfuscate the identification of people included in the audience for a message. The email in Figure 3.6 shows a message which illustrates the three classes of email recipients that are commonly accepted within email conversations: to, carbon copy ('Cc') and blind carbon copy ('Bcc') recipients. Note that only the sender is able to see the inclusion of any 'Bcc' recipients; the message leaves no evidence for other recipients that a 'Bcc' recipient has been included.

In Figure 3.6, the sole 'To' recipient is Laura Johnson. The message also has ten 'Cc' recipients, and a single 'Bcc' recipient. By design, the 'To' field recipients are the primary audience of the message, 'Cc' field recipients are others whom the sender wishes to publicly inform of the message, and 'Bcc' field recipients are those surreptitiously being informed of the communication. It would be convenient if, for the purposes of identifying obligation acts, we could assume that actionable content would be addressed only to 'To' recipients. In practice, however, this neat separation of the different roles of various participants often fails to reflect the actual participant roles. We frequently see, for example, obligation acts on which a carbon-copied recipient is required to act.

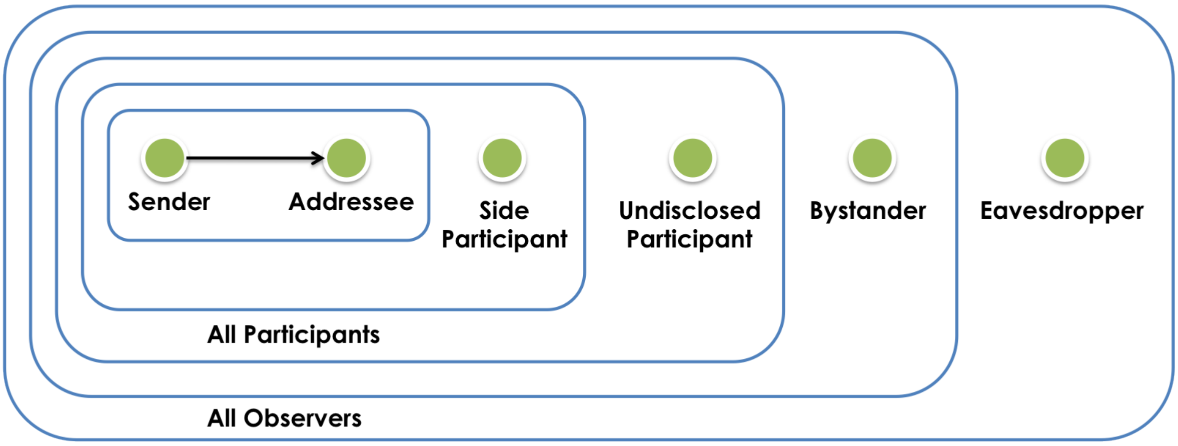


Figure 3.7: Different roles within the audience for an email message, adapted from (Clark, 1997, p14).

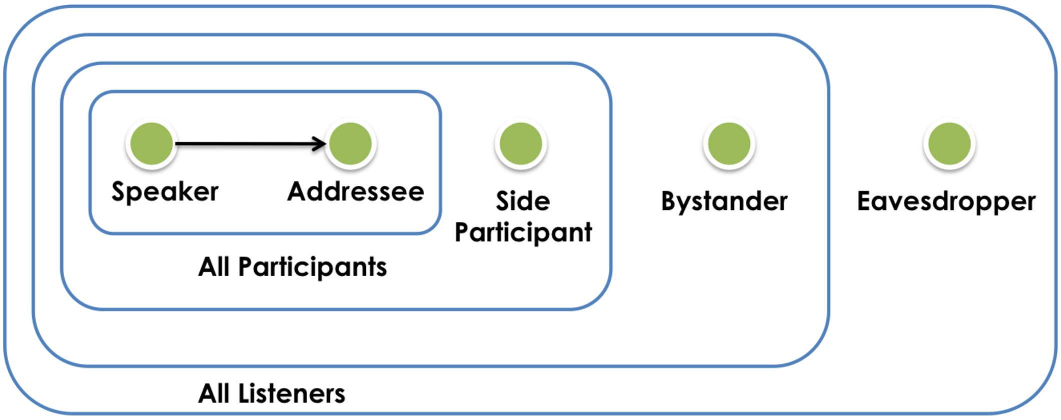


Figure 3.8: Different roles within the audience for a spoken utterance (Clark, 1997, p14).

Figure 3.7 presents a model for the various audience roles in an email conversation. This is an adaptation of a model proposed by Clark (1997, p. 14) for spoken conversation that is shown in Figure 3.8. The primary difference is the inclusion of an additional class of Undisclosed participants who can be part of the audience in an email conversation. We justify this modification below.

The primary participants in an email conversation are the sender and recipients. For any utterance in an email message, there are one or more addressees. An addressee may be explicitly identified, as in the utterance *Finn, please send me your comments before 4pm today*, or may be implicitly associated with an utterance through the use of an opening message greeting (e.g., *Dear Finn and Ronan*), or simply from the list of message recipients identified in the email headers. In an email message, the set of addressees may vary for different utterances in the same message, in the same way that spoken utterances may be directed at different participants in a group during the course of a conversation. Email messages can also obscure the identity of the recipients because an email address may be unique to a person or may represent a mailing list or other group, the membership of which is not always revealed. Where an email is sent to a list such as `managers@enron.com`, for example, it can be unclear to any single recipient exactly who else is being addressed by the content in that message, even though the address itself is visible.

In addition to the implicit or explicit addressees, there may also be other side participants. In a canonical email conversation, the side participants would be the group of carbon copy recipients. In practice, however, the set of side participants for any particular utterance includes any email recipients, whether to or cc recipients, who are not being directly addressed by that utterance.

In a departure from Clark's original model of conversational participation, we introduce an additional class of participant: the undisclosed participant. An undisclosed participant is someone in the set of blind carbon copy recipients. These recipients are generally known only to the email sender. Neither the public recipients, nor any other 'Bcc' recipients of the message, are aware that these participants are observing the conversation. Undisclosed participants differ from Clark's bystanders in a number of important ways. Firstly, undisclosed recipients are participants in the conversation, unlike bystanders, since the sender is aware of their presence, and has chosen to address his or her message to them. Also unlike bystanders, undisclosed participants are not openly present; their presence in the conversation is not public to any of the participants, other than the sender.

Despite the conversation itself happening in a virtual space, the email users who are participating are inevitably situated in physical space. There thus exists a similar opportunity for bystanders, overhearers who are openly present for the sender or recipient but are not part of the conversation. In an email conversation, bystanders would primarily be people who are in the same physical space as the email user, such as colleague in the user's cubicle or office space, who can see the user's email displayed on the screen.

The final category of eavesdroppers includes people who are not participants, but observe the email conversation without the sender's or any of the recipients' awareness. As a written medium, eavesdroppers to an email conversation (and other varieties of overhearers noted by Clark) may include IT staff monitoring email communication within company networks, or law enforcement and intelligence organisations performing targeted interception of email communication.

For the purposes of interpreting requests and commitments in a message, we focus on the addressees and side participants, and occasionally on undisclosed participants. We ignore the possible existence of bystanders and eavesdroppers, as any requests or commitments in a message are unlikely to be intentionally addressed to such an audience.

3.1.5 Adopting a Specific Perspective

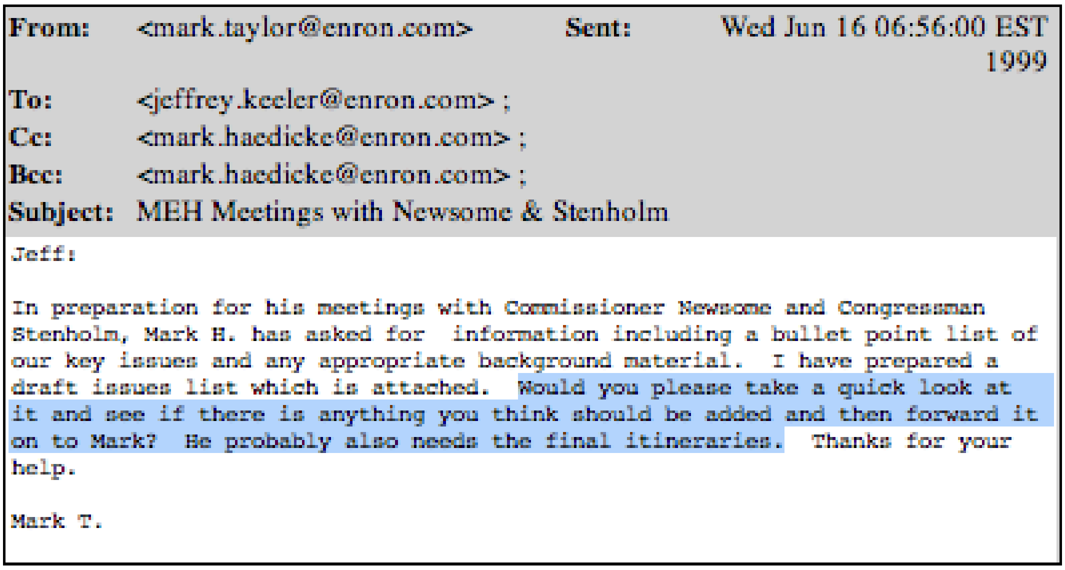
As we saw in Examples (3.54) and (3.55), the same utterance can simultaneously convey different speech acts to different audiences. Discerning whom the sender intends to address with an utterance can be informed by both the audience of a message, as discussed in Section 3.1.4, and by the specific ADDRESSIVITY of a message or utterance. We use the term addressivity to refer to whom a particular utterance, collection of utterances (e.g., a paragraph), or message is directed towards. In the case of a request, this also identifies who the AGENT of the associated action is. Because the interpretation of a request or a commitment depends on whose perspective is adopted, variations in addressivity add further complexity to identifying requests and commitments. In this section, we consider how to identify the intended audience for a message or an utterance, and how this can affect interpretation of specific utterances.

Terms of address are one way of identifying the intended audience of a message or utterance. Terms of address frequently occur at the message level, generally in the form of a message greeting. With or without a message greeting, specific request utterances can also be addressed to different people or groups, as in Example (3.56).

(3.56) *Finn, please don't forget to send me your thoughts.*

When we ask annotators to analyse utterances in an email message, we instruct them to consider the email text from the point-of-view of a single, specific recipient, drawn from the set of all identifiable recipients of the message. In our search of the literature around identifying speech acts in email, we have found that this framing of whose perspective is adopted when interpreting requests and commitments is not explicitly discussed. It is, however, important, since whether a particular utterance (or message) functions as a request or commitment can be different for different recipients. The example email in Figure 3.9 demonstrates this point. Although both Mark Haedicke and Jeffrey Keeler are recipients of the message, the content only contains a request for Jeffrey, not for Mark. Other messages can contain different sets of requests or commitments for different recipients. We have seen one example of an utterance that acts as a request for some recipients and commitment for others in Examples (3.54) and (3.55). Another example is shown in Figure 3.10, where different tasks are explicitly addressed to different recipients. If we were adopting Mitch Robinson's point-of-view as our specific perspective for identifying obligation acts, then we would not identify the utterance addressed to Richard (*Richard, any ideas or comments?*), since this does not represent a request for Mitch; it only functions as a request for Richard.

We also recognise that messages and utterances can legitimately function simultaneously as both requests and commitments from the perspective of a single recipient. For example, the utterance *Let me know if I can help with editing the document* in an email from Ronan to Finn should be identified as both a request—'please respond to me'—and a commitment—'I will help to edit the document if you accept my offer'. We show both these interpretations represented below:



From:	<mark.taylor@enron.com>	**Sent:**	Wed Jun 16 06:56:00 EST 1999
To:	<jeffrey.keeler@enron.com> ;		
Cc:	<mark.haedicke@enron.com> ;		
Bcc:	<mark.haedicke@enron.com> ;		
Subject:	MEH Meetings with Newsome & Stenholm		

```
Jeff:

In preparation for his meetings with Commissioner Newsome and Congressman
Stenholm, Mark H. has asked for  information including a bullet point list of
our key issues and any appropriate background material.  I have prepared a
draft issues list which is attached.  Would you please take a quick look at
it and see if there is anything you think should be added and then forward it
on to Mark?  He probably also needs the final itineraries.  Thanks for your
help.

Mark T.
```

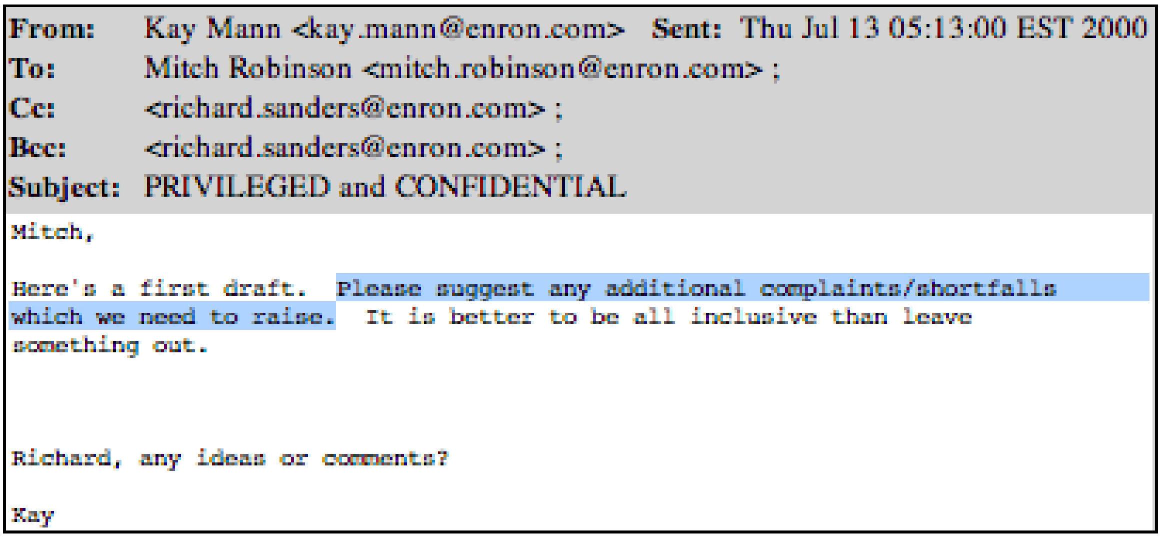
FIGURE 3.9: An email illustrating that requests and commitments may differ between recipients of the same message.

(3.57) R = ⟨ACTION=RESPOND, REQUESTOR=RONAN, REQUESTEE=FINN⟩
UR = ⟨DIRECT, *let me know if I can help with editing the document*, R⟩
C = ⟨ACTION=EDIT DOCUMENT, COMMITTOR=RONAN,
COMMITTEE=RONAN⟩
UC = ⟨INDIRECT, *let me know if I can help with editing the document*, C⟩

The utterance in Example (3.57) is considered an OFFER. This duality of interpretation is particularly common for offers, which by their very nature are often commitments conditioned on a response, request or acceptance from the recipient. We talk more about conditional requests and commitments in Section 3.1.7, and focus specifically on analysing offers in Section 3.3.3.1.

3.1.6 Focusing on New Content from the Current Author

Content in an email message may be new content from the current sender, or may represent content that comes from previous messages in an ongoing email conversation. Because email is a written medium, a sender may, for example, forward another sender's message to an audience entirely different from its original or intended audience. When identifying task-related content, such usage can be ambiguous as to whether requests and commitments in the original (forwarded) message are actually now directed at the new recipient, intending to make them an active participant in the original conversation, or whether the current recipients of the forwarded message are simply being added as



From: Kay Mann <kay.mann@enron.com> **Sent:** Thu Jul 13 05:13:00 EST 2000
To: Mitch Robinson <mitch.robinson@enron.com> ;
Cc: <richard.sanders@enron.com> ;
Bcc: <richard.sanders@enron.com> ;
Subject: PRIVILEGED and CONFIDENTIAL

```
Mitch,

Here's a first draft.  Please suggest any additional complaints/shortfalls
which we need to raise.  It is better to be all inclusive than leave
something out.



Richard, any ideas or comments?

Kay
```

Figure 3.10: An email illustrating that requests and commitments may differ between recipients of the same message.

side participants, bystanders or eavesdroppers. This adds complexity to the process of identifying requests and commitments.

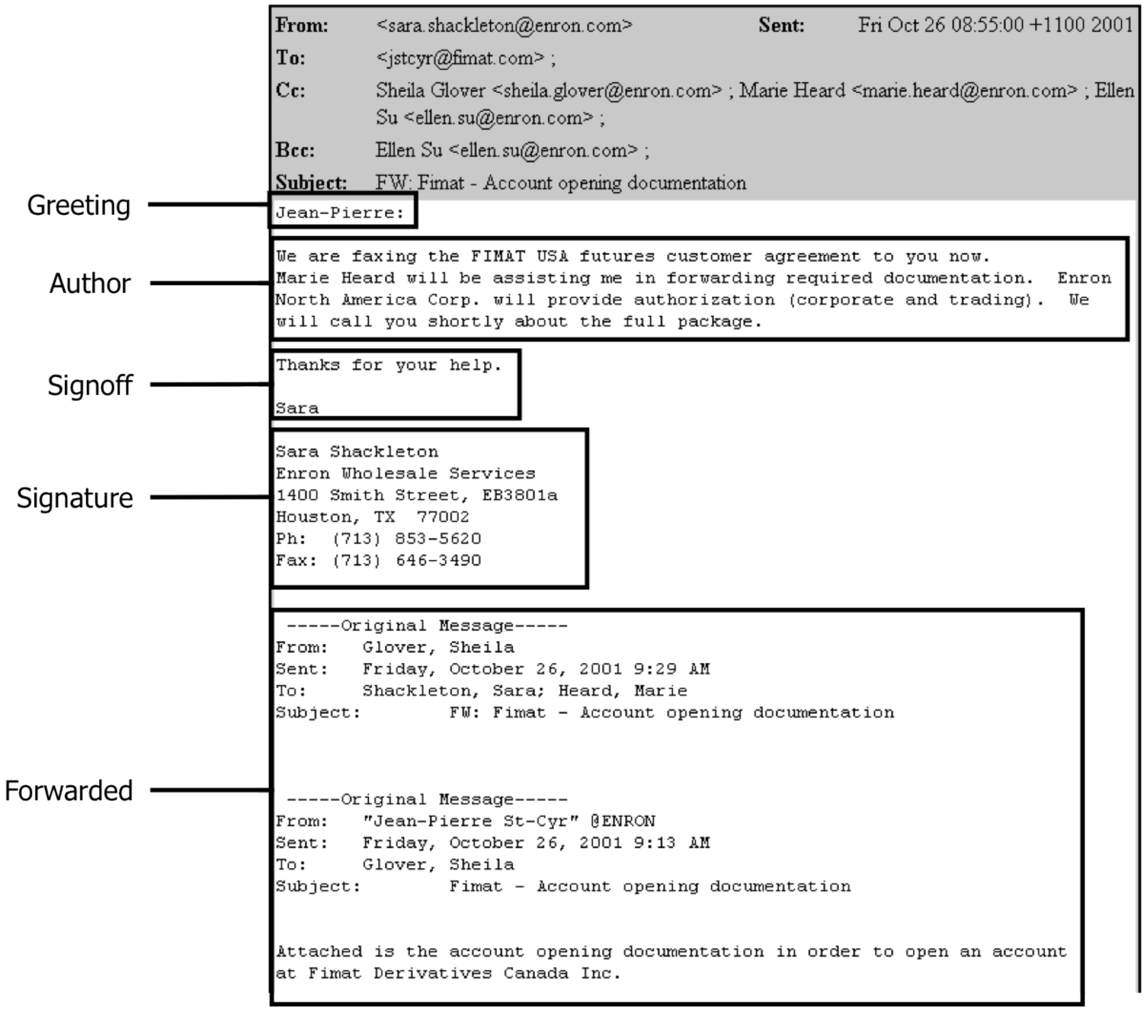
An email message usually has several different functional parts, which we call email zones, as illustrated in Figure 3.11. We instruct our annotators to only identify requests and commitments within the specific zones that contain new content authored by the current message sender that are directed at the identified recipient of the message. We do not seek to identify any requests or commitments which appear in text that is directed at another agent, or in the content quoted from earlier messages, or in other email zones that do not contain content authored by the current sender, such as email signatures, automatically inserted advertising or legal disclaimers. We aim to use such material only to interpret or clarify the intent of new content in an email message. In practice, in our annotation experiments, annotators exercised their own judgement, and occasionally marked requests or commitments in forwarded content where they interpreted such content as conveying an obligation for the specified email recipient, despite the content being authored by a third-party.

We discuss how we identify these different email zones within a message during automatic request and commitment classification in Section 4.4.

3.1.7 Dealing with Conditional Requests and Commitments

Consider the following request from Tim Belden to a group of recipients from the perspective of one of the recipients, Robert Badeer:

(3.58) R = ⟨Action=Contact Debra, Requestor=Tim Belden,

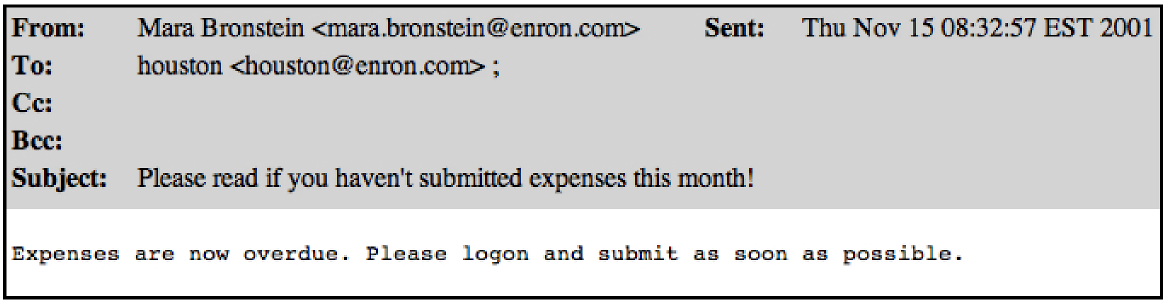


```
From:     <sara.shackleton@enron.com>               Sent:   Fri Oct 26 08:55:00 +1100 2001
To:       <jstcyr@fimat.com> ;
Cc:       Sheila Glover <sheila.glover@enron.com> ; Marie Heard <marie.heard@enron.com> ; Ellen
          Su <ellen.su@enron.com> ;
Bcc:      Ellen Su <ellen.su@enron.com> ;
Subject:  FW: Fimat - Account opening documentation

Jean-Pierre:

We are faxing the FIMAT USA futures customer agreement to you now.
Marie Heard will be assisting me in forwarding required documentation.  Enron
North America Corp. will provide authorization (corporate and trading).  We
will call you shortly about the full package.

Thanks for your help.

Sara

Sara Shackleton
Enron Wholesale Services
1400 Smith Street, EB3801a
Houston, TX  77002
Ph:  (713) 853-5620
Fax: (713) 646-3490

 -----Original Message-----
From:   Glover, Sheila
Sent:   Friday, October 26, 2001 9:29 AM
To:     Shackleton, Sara; Heard, Marie
Subject:        FW: Fimat - Account opening documentation



 -----Original Message-----
From:   "Jean-Pierre St-Cyr" @ENRON
Sent:   Friday, October 26, 2001 9:13 AM
To:     Glover, Sheila
Subject:        Fimat - Account opening documentation


Attached is the account opening documentation in order to open an account
at Fimat Derivatives Canada Inc.
```

Figure 3.11: Different email zones in an email message.

Requestee=Robert Badeer⟩
UR = ⟨Direct, *please let Debra know if you have a scheduling conflict and cannot attend*, R⟩

This request demonstrates an obligation to respond that is only created for Robert Badeer, as a recipient of the message, if the initial condition (having a scheduling conflict) holds true. We call such requests conditional requests. Conditional requests are very common in the workplace email communication we have examined. The conditionality can involve stated or implied actions, as well as specific states or other criteria that are required to occur or be satisfied before the obligation for a recipient in a request is triggered. The stated condition is often included in the same utterance as the request, but may instead be elsewhere in the email message, such as in the subject line, as shown in Figure 3.12.



From:	Mara Bronstein <mara.bronstein@enron.com>	**Sent:**	Thu Nov 15 08:32:57 EST 2001
To:	houston <houston@enron.com> ;		
Cc:			
Bcc:			
Subject:	Please read if you haven't submitted expenses this month!		

```
Expenses are now overdue. Please logon and submit as soon as possible.
```

Figure 3.12: An example of a message with conditionality defined in the subject line which has scope over the entire message, including the request in the second sentence.

To adequately capture conditional requests, we add an additional conditional attribute to our ontological representation. In doing so, it is important to note that the conditionality of a request or commitment is a property of the underlying act itself, rather than of a specific realisation of the act. For this reason, we capture conditionality in the representation of the request or commitment. Below we augment the representation of Example (3.58) to capture the conditionality:

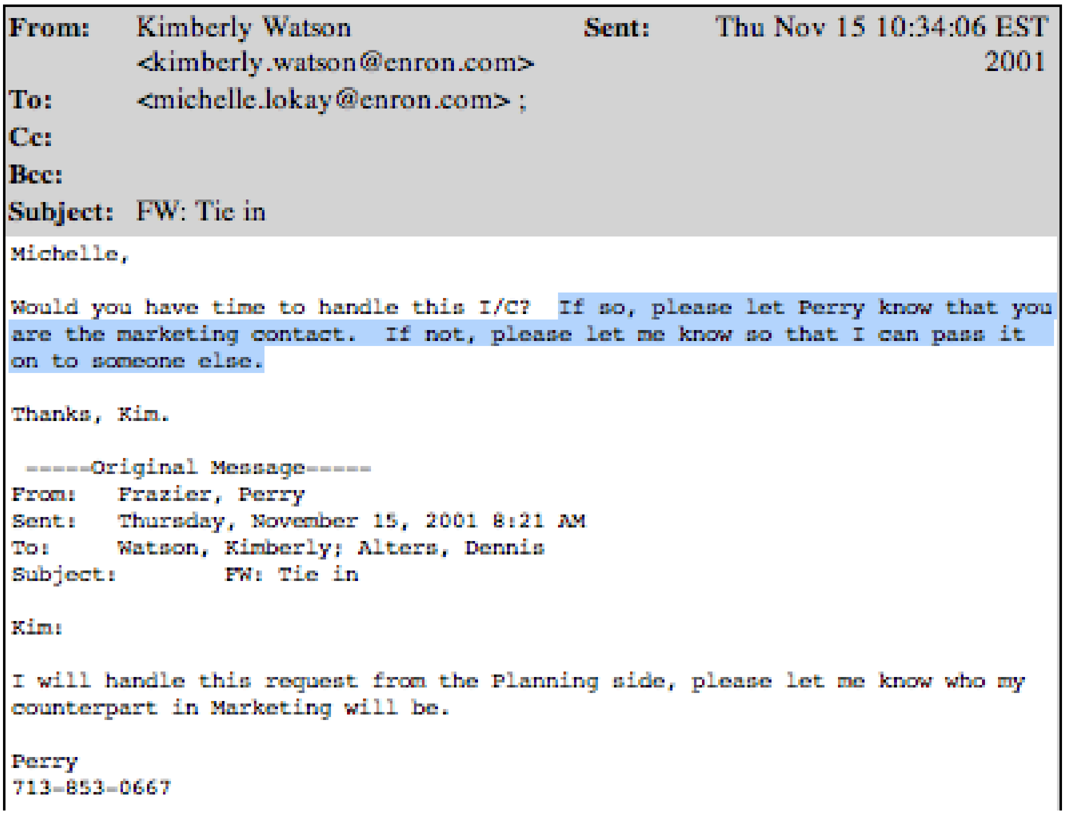
(3.59) R = ⟨Action=Contact Debra, Requestor=Tim Belden,
Requestee=Robert Badeer,
Condition=Scheduling Conflict/Cannot Attend ⟩

Another example of a conditional request is shown in the email in Figure 3.13. This message has two separate conditional requests from Kimberly highlighted, both of which are extracted below:

(3.60) UR = ⟨Direct, *if so, please let Perry know that you are the marketing contact*, R ⟩
R = ⟨ Action=Contact Perry, Requestor=Kimberly Watson,
Requestee=Michelle Lokay,
Condition=Have time to handle ⟩

(3.61) UR = ⟨Direct, *if not, please let me know so that I can pass it on to someone else*, R ⟩
R = ⟨ Action=Contact Kimberly, Requestor=Kimberly Watson,
Requestee=Michelle Lokay,
Condition=No time to handle ⟩

Given that the conditions applied to these two requests are mutually exclusive—either Michelle has the time to handle the request, or she does not—only one of the two requests will be actionable for Michelle.



From: Kimberly Watson <kimberly.watson@enron.com> **Sent:** Thu Nov 15 10:34:06 EST 2001
To: <michelle.lokay@enron.com> ;
Cc:
Bcc:
Subject: FW: Tie in

```
Michelle,

Would you have time to handle this I/C?  If so, please let Perry know that you
are the marketing contact.  If not, please let me know so that I can pass it
on to someone else.

Thanks, Kim.

 -----Original Message-----
From:   Frazier, Perry
Sent:   Thursday, November 15, 2001 8:21 AM
To:     Watson, Kimberly; Alters, Dennis
Subject:        FW: Tie in

Kim:

I will handle this request from the Planning side, please let me know who my
counterpart in Marketing will be.

Perry
713-853-0667
```

Figure 3.13: An email containing two mutually-exclusive conditional requests.

Commitments may similarly be conditional, and we refer to such utterances as CONDITIONAL COMMITMENTS. An example of a conditional commitment is the utterance below extracted from an email sent from Mara Bronstein to Kim Ward:

(3.62) C = ⟨ ACTION=PRINT DOCUMENT, COMMITTOR=MARA BRONSTEIN,
COMMITTEE=MARA BRONSTEIN,
CONDITION=ASK MARA TO PRINT ⟩
UC = ⟨DIRECT, *if you would rather me just print this out for you, I will*, C ⟩

In Example (3.62), Mara is only committed to the future action of printing the document if Kim, the recipient, requests her to complete the action. This particular example is an OFFER, a common form of conditional commitment that we discuss further in Section 3.3.3.1.

Conditional commitments are frequently realised as offers, although they can be conveyed by a range of conditional commissive acts, including:

(3.63) Offer: *I'll finish it for you if you don't finish that work today.*

(3.64) Threat: *I'll fire you if you don't finish that work today.*[8]

(3.65) Conditional Promise: *If you finish the document early, I promise you can go home early.*

(3.66) Conditional Invitation: *If you have not already done so, please visit our Finance Centre for more information.*

Utterances may also simultaneously convey both a conditional commitment and a request. Both our own observation of annotated data, and previous linguistic work (Hancher, 1979), suggest that conditional commitments frequently join a request and a commitment together. Example (3.67) shows such an utterance, drawn from an email sent from Rick Buy to Dave Thomas that conveys both a conditional commitment and a conditional request.

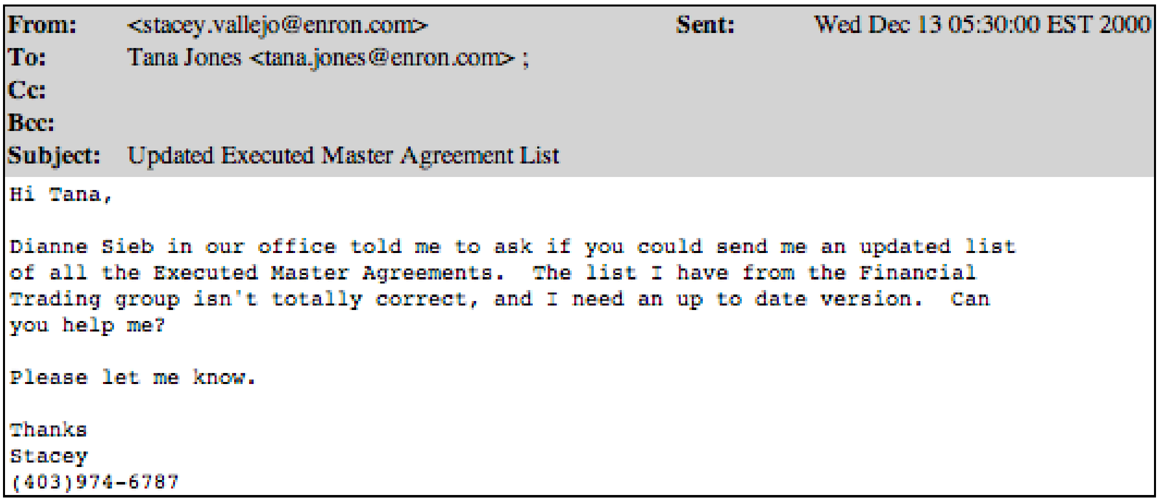
(3.67) C = ⟨ Action=Put in touch with Risk Assessment representative,
Committor=Rick Buy, Committee=Cassandra Schultz,
Condition=Question asked ⟩
UC = ⟨Direct, *if you have any questions, please contact Cassandra Schultz who will put you in touch with the appropriate Risk Assessment & Control representative*, C ⟩

R = ⟨ Action=Contact Cassandra Schultz, Requestor=Rick Buy,
Requestee=Dave Thomas,
Condition=Have question ⟩
UR = ⟨Direct, *if you have any questions, please contact Cassandra Schultz who will put you in touch with the appropriate Risk Assessment & Control representative*, R ⟩

As we discuss in Section 3.2, our experiments in asking people to identify requests and commitments in real-world email messages suggest that there is some contention about whether utterances such as Example (3.67) contain only a request for the recipient, or whether they also convey a commitment from the speaker to follow through with the stated action if the recipient accepts or acts on their offer. In the specific case of Example (3.67), the explicit nature of the commitment suggests that this utterance conveys both a conditional request (*please contact Cassandra if you have any questions*) and an offer (*Cassandra will put you in touch with the appropriate people if you contact her*).

In definitional terms, both conditional requests and conditional commitments are valid forms of obligation acts, given our goal of identifying all requests and commitments present in an email message that are implicitly or explicitly addressed to the specified recipient.

[8] As we discuss in Section 3.3.3.3, we do not consider threats to be commitments for our work in this thesis.



From:	<stacey.vallejo@enron.com>	**Sent:**	Wed Dec 13 05:30:00 EST 2000
To:	Tana Jones <tana.jones@enron.com> ;		
Cc:			
Bcc:			
Subject:	Updated Executed Master Agreement List		

```
Hi Tana,

Dianne Sieb in our office told me to ask if you could send me an updated list
of all the Executed Master Agreements.  The list I have from the Financial
Trading group isn't totally correct, and I need an up to date version.  Can
you help me?

Please let me know.

Thanks
Stacey
(403)974-6787
```

Figure 3.14: An email message with numerous request utterances. The set of obligation utterances that should be identified depends on the end application.

3.1.8 Considering the Purpose and Relevance

As a final consideration in our definitions, we take account of the *purpose* of identifying and extracting requests and commitments, in the sense of considering the end application or intended use of the identified requests and commitments. Intended use is a particularly important consideration, since different sets of uttered requests and commitments from the same email message would be considered relevant for different end uses and purposes.

When a user is reading an email message with access to the entire message content, for example, the uttered requests and uttered commitments that users would like to have highlighted are likely to differ from those that should be extracted and aggregated into a separate task list or task dashboard, where the requests and commitments will be read and interpreted in isolation from their original message content and context. The set of tasks to include in an action-based summary of a message or thread, displayed in the inbox alongside existing message metadata (such as sender, date, subject and so on), would be different again.

As a simple example, consider a request for inaction, such as:

(3.68) *Please do not distribute the attached draft document.*

It might well be relevant to highlight the uttered request in Example (3.68) within the message body to help focus the attention of a user browsing the complete message content to the actionable content and obligations. In a different context, however, where we are identifying requests to extract to a user's task list, this same request would not be relevant to identify. We discuss requests and commitments for inaction in more detail in Section 3.3.6.

The issue is broader than requests for inaction. Consider the email in Figure 3.14. If our goal is to highlight requests and commitment *in situ*, then it might be relevant to identify all the following uttered requests and commitments:

(3.69) *Dianne Sieb in our office told me to ask if you could send me an updated list of all the Executed Master Agreements.*

(3.70) *I need an up to date version.*

(3.71) *Can you help me?*

(3.72) *Please let me know.*

If, however, our goal is to extract obligations to a standalone task list, we would focus on extracting or synthesising a single representative task entry for each distinct, underlying request and commitment. We would not create a separate task entry for each *uttered* request or commitment, since they frequently restate the same underlying obligation. In Figure 3.14, for example, we might add a single request to the user's task list, which captured the underlying request to provide Stacey with an updated list of all the Executed Master Agreements. The exact wording of the extracted task could be that of a representative uttered request, or might involving combining, expanding or re-writing a series of related uttered requests or commitments to capture the required context so that the task can make sense separate from the message context from which it was drawn. In this case, for example, we might re-write the extracted request to something like *Send Stacey an updated list of all the Executed Master Agreements.*

When manually or automatically identifying obligation acts, we aim to identify all requests and commitments that would be useful to highlight for a recipient while they are reading an email message. We thus attempt to identify all realisations of every request and commitment. We consider the task of selecting from these identified utterances for specific applications as a separate and subsequent task that can be tackled with approaches such as ranking or filtering of the obligation act result set for each message. We do not address this ranking or filtering task in this thesis; instead we leave this challenge to future work.

3.1.9 A Summary of the Scope of Requests and Commitments

In this section, we have specified definitions for identifying both requests and commitments in workplace email. Requests are utterances that place an obligation for current of future action on one or more email recipients. Commitments are utterances which place an obligation for current or future action on the email sender, or some other party. We have outlined the possible participants in an email conversation, and discussed the importance of adopting a specific perspective and purpose, as well as the need to consider the functional structure of the message, to ensure only relevant requests and commitments are identified. To support a deeper analysis of these obligation acts in Section 3.3, we have also presented a formal representation that captures the agents and communicators of each act, along with detail about any conditions that must be met for the obligation to hold, and detail of the specific utterance that was used to realise the act.

3.2 Creating Annotated Email Corpora

The definitions and analysis that we presented in Section 3.1 draw on both theoretical work in Speech Act Theory, and on empirical observations of real-world request and commitment use in workplace email. To facilitate this empirical understanding of how people use requests and commitments, we have gathered independent human interpretations of obligation acts within messages drawn from the Enron email corpus. The resulting annotated datasets provide the empirical basis on which we have built our definitions and our understanding of requests and commitments. The analysis of this data has allowed us to better understand how people make requests and commitments in real-world email communication, and has clearly demonstrated that requests and commitments are more complicated than a purely theoretical treatment might lead one to believe.

In this section, we describe both the Enron email corpus that provides our source of real-world email messages, and the annotation experiments that we have undertaken to enrich data from the Enron corpus with pragmatic interpretations and annotations. We present an overview of our annotation experiments, the task guidelines we provided to our human annotators, the resulting inter-annotator agreements between our annotators, and some analysis of the major sources of disagreement for each experiment. We then analyse and reconcile common cases of subtle ambiguity and complexity as specific functional features of obligation acts in Section 3.3. More detail on the guidelines and experiments is included in Appendix A and Appendix B respectively.

3.2.1 The Enron Email Corpus

In the early 2000s, a large corpus of real-world email messages was subpoenaed from Enron Corporation and placed on the public record during a US Federal Energy Regulatory Commission (FERC) investigation (Klimt and Yang, 2004). Thanks to the work by several academics, an electronic version of this corpus was made available to researchers and has since become a widely used data source for email research. There are several different versions and formats of the corpus available. The raw data of the Enron corpus consists of over 500,000 email messages from the email accounts of approximately 150 senior managers at Enron over approximately an 18 month period.

We employ the database dump of the corpus curated and released by Andrew Fiore and Jeff Heer.[9] This version of the corpus has been processed in an attempt to remove duplicate email messages and to normalise sender and recipient names, resulting in a collection containing just over 250,000 email messages. The content of these messages contains almost 12 million lines of text and more than 3 million sentences.[10] Within this

[9] This corpus is available from `http://bailando.sims.berkeley.edu/enron/enron.sql.gz`.

[10] These metrics are derived using our own line and sentence-splitting tools over the Fiore and Heer version of the Enron corpus.

collection, there are more than 20,000 distinct senders, and more than 78,000 different email recipients. This version of the corpus does not include attached documents, which creates some additional ambiguity for classes of requests that involved acting on content in an attached document, but is otherwise inconsequential, given our focus on the language in the actual email messages, rather than in business documents more generally. In all our annotation experiments, we annotate text from the message body and subject line content of emails extracted from this corpus.

Each of the annotation tasks we describe in this chapter was undertaken using real-world email messages that are sampled from Fiore and Heer's version of the Enron Email Corpus.

3.2.2 A Summary of Our Email Annotation Experiments

We conducted a series of seven separate annotation tasks in which independent human annotators explicitly identified requests and commitments in email messages, drawn from the Enron dataset, at different levels of granularity:

- Message level: applying annotation judgements to an entire email message;
- Sentence level: applying annotation judgements to each sentence from the content of an email message; and
- Span level: identifying relevant spans of text of variable length from an email message, and applying annotation judgements to these spans.

In aggregate, across these experiments we gathered two or more independent sets of annotation for more than 2750 distinct email messages. This collection of annotated data facilitates our analysis of real-world requests and commitments in this chapter. The inter-annotator agreement levels that we present in the next section demonstrate that there are areas of disagreement which belie the underlying complexity in the way requests and commitments are realised and used in workplace email communication. The specific instances of disagreement provide data to support our analysis of the complex functional phenomena that we present in Section 3.3.

Overall, however, despite the observed complexity, our annotation experiments demonstrate levels of agreement that are sufficient for the generated labelled data to be used to train and evaluate the automated request and commitment classifiers that we present in Chapter 4 and Chapter 5.

Below we briefly present our series of annotation experiments, along with a synopsis of issues discovered along the way, and changes we made to our annotation guidelines in response. Further details of each annotation experiment is included in Appendix B.

3.2.3 Our Fixed Unit Annotation Tasks

Table 3.1 shows a high-level summary of inter-annotator agreement scores from the first five of our manual annotation experiments, which were all fixed unit annotation tasks, focused on identifying of requests and commitments at the message and sentence

				3-way κ	
Experiment	Annotators	Msgs	Sentences	Request	Commit
Sentence 1	3	54	310	0.78	0.54
Sentence 2	3	350	750	0.80	0.74
Message 1	3	100	–	0.84	0.78
Message 2	5	209	–	0.79	0.80 [12]
Message 3	3	664	–	0.85	0.86

Table 3.1: Three-way κ agreement from fixed-unit email annotation experiments.

levels. Details of how messages were selected for each experiment, and the annotation guidelines that were provided to annotators, are described in detail in Appendix B. In each case, messages and sentences were independently marked by between three and five different people. Agreements are reported as three-way Cohen's κ scores, which represent pair-wise κ scores averaged across each pair of annotators. These agreements refer to binary agreement about whether a specific sentence or message contains one or more requests or commitments. The first two experiments required annotators to identify requests and commitments at the sentence level,[11] while the third and fourth experiments involved annotations at the message level. In both cases, the unit of annotation was fixed as either a single sentence or the entire message, so the annotator agreement reported refers to the labels applied to these fixed units of text.

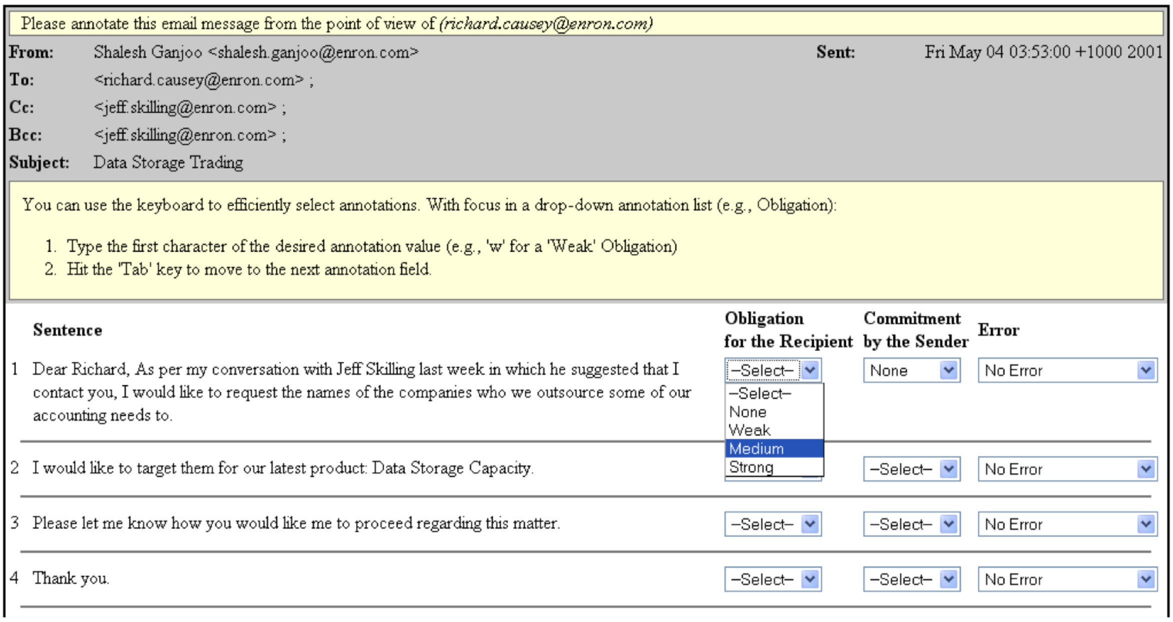
Our first annotation experiment (*Sentence 1* in Table 3.1) is described in detail in Section B.1. We used the annotation guidelines included in Section B.1.4, and had annotators mark real-world email messages using a web-based annotation tool shown in Figure 3.15. This annotation tool required annotators to provide two judgements for each sentence in each annotated message:

1. Are there any requests, and if so, what strength of obligation does it represent?
2. Are there any commitments, and if so, what strength of obligation does it represent?

This first annotation experiment was the only time that we asked annotators to indicate the strength of obligation for requests and commitments, as one of *weak*, *medium* or *strong* for each obligation act that was marked. The definitions for each of these levels of strength are included in Section B.1.4. Identifying the strength of an obligation act turned out to be a highly subjective task, and one that was impossible to

[11] We use the SentParBreak sentence and paragraph segmenter (Piao, Wilson, and McEnery, 2002), augmented with our own email specific preprocessing, in order to identify sentences to annotate in our sentence annotation tasks.

[12] These results represent the best three-way agreement across our five annotators.



Please annotate this email message from the point of view of *(richard.causey@enron.com)*

From: Shalesh Ganjoo <shalesh.ganjoo@enron.com> **Sent:** Fri May 04 03:53:00 +1000 2001
To: <richard.causey@enron.com> ;
Cc: <jeff.skilling@enron.com> ;
Bcc: <jeff.skilling@enron.com> ;
Subject: Data Storage Trading

You can use the keyboard to efficiently select annotations. With focus in a drop-down annotation list (e.g., Obligation):

1. Type the first character of the desired annotation value (e.g., 'w' for a 'Weak' Obligation)
2. Hit the 'Tab' key to move to the next annotation field.

	Sentence	Obligation for the Recipient	Commitment by the Sender	Error
1	Dear Richard, As per my conversation with Jeff Skilling last week in which he suggested that I contact you, I would like to request the names of the companies who we outsource some of our accounting needs to.	–Select–	None	No Error
2	I would like to target them for our latest product: Data Storage Capacity.		–Select–	No Error
3	Please let me know how you would like me to proceed regarding this matter.	–Select–	–Select–	No Error
4	Thank you.	–Select–	–Select–	No Error

FIGURE 3.15: The web-based annotation tool we developed and used in the first sentence-level annotation task.

do systematically without having better access to the context around each message. For these reasons, our remaining annotation tasks focus only on the identification of obligation acts, without asking annotators to rank their strength.

In addition to marking requests and commitments, we asked each annotator to indicate whether there were any processing errors, such as incorrect sentence segmentation or invalid character encodings. This was done to provide us with some quantitative feedback on the error rates of our sentence segmentation, message filtering and data extraction processes and tools.

From the annotations of our three independent annotators who took part in this initial experiment, we drew the following tentative conclusions about human agreement in identifying requests and commitments in email:

1. There is good agreement ($\kappa = 0.78$) about which sentences embody a request.

2. There is some tentative agreement ($\kappa = 0.60$) about the strength of requests.

3. There is poorer agreement about which sentences embody a commitment ($\kappa = 0.54$) and poor agreement about the strength of those commitments ($\kappa = 0.37$).

A systematic source of disagreement between annotators concerned whether to classify a sentence that embodies an OFFER as a commitment. An example from this pilot annotation task is shown in Example (3.73).

(3.73) *If you have any questions, please contact Cassandra Schultz who will put you in touch with the appropriate Risk Assessment & Control representative*

Our annotators disagreed as to whether the utterance in Example (3.73) commits Cassandra Schultz to some future action. Some annotators classified this sentence as containing only a request for the recipient, while others judged that it contained both a request for the recipient and a commitment for Cassandra, who implicitly promises to make good on their offer if the recipient accepts or acts on it. Almost 40% of disagreements from this initial experiment stemmed from similar disagreements as to whether offers are considered to represent commitments.

Another common source of disagreement stemmed from interpretations of implicit requests. In our pilot annotation corpus, these occur frequently as statements about meetings. Consider, for example, the sentence:

(3.74) *Ken wants to have a meeting this afternoon in regard to California from a PR standpoint.*

Our annotators disagreed as to whether the sentence in Example (3.74) represents a request to attend a meeting. In the original email this sentence is followed by this explicitly directive sentence:

(3.75) *Please let me know if you will be able to attend*

Example (3.75) is clearly a request, but its presence creates some confusion over the interpretation of Example (3.74). In the absence of Example (3.75), the sentence in Example (3.74) should, we would argue, be interpreted as an indirect request for the recipient. When followed by Example (3.75), however, this interpretation as an indirect request is now redundant, given the explicit restating of the request that follows. For our pilot annotation task, we originally directed annotators to annotate each sentence "in the context of the entire email message", which created some confusion around whether the first sentence should not be annotated as a request. It became clear to us that our annotation guidelines needed to provide more specific guidance to annotators about how to interpret such situations. From this stemmed the guidance we discussed in Section 3.1.8, around marking *all* uttered requests and commitments. In this example, it would mean that annotators should identify both Example (3.74) and Example (3.75) as requests.

Another issue noted by some of our annotators was that differences in the relationship between the sender and the specified recipient (specifically, differences in the 'importance' of the sender) would affect the strength they associated with requests or commitments. A specific example on which annotators commented is:

(3.76) *You might like to organise the paper from a broad overview of the electrical market in the west (including basic descriptions, timelines, fundamentals etc.) down to a specific description of what you did at Enron*

Two annotators commented that their interpretation of the strength of any request would be different depending on the organisational relationship between sender and receiver. If this represented, for example, a manager assigning a task to a direct report, then the sentence would be interpreted to have a different strength than if it were a request from a peer.

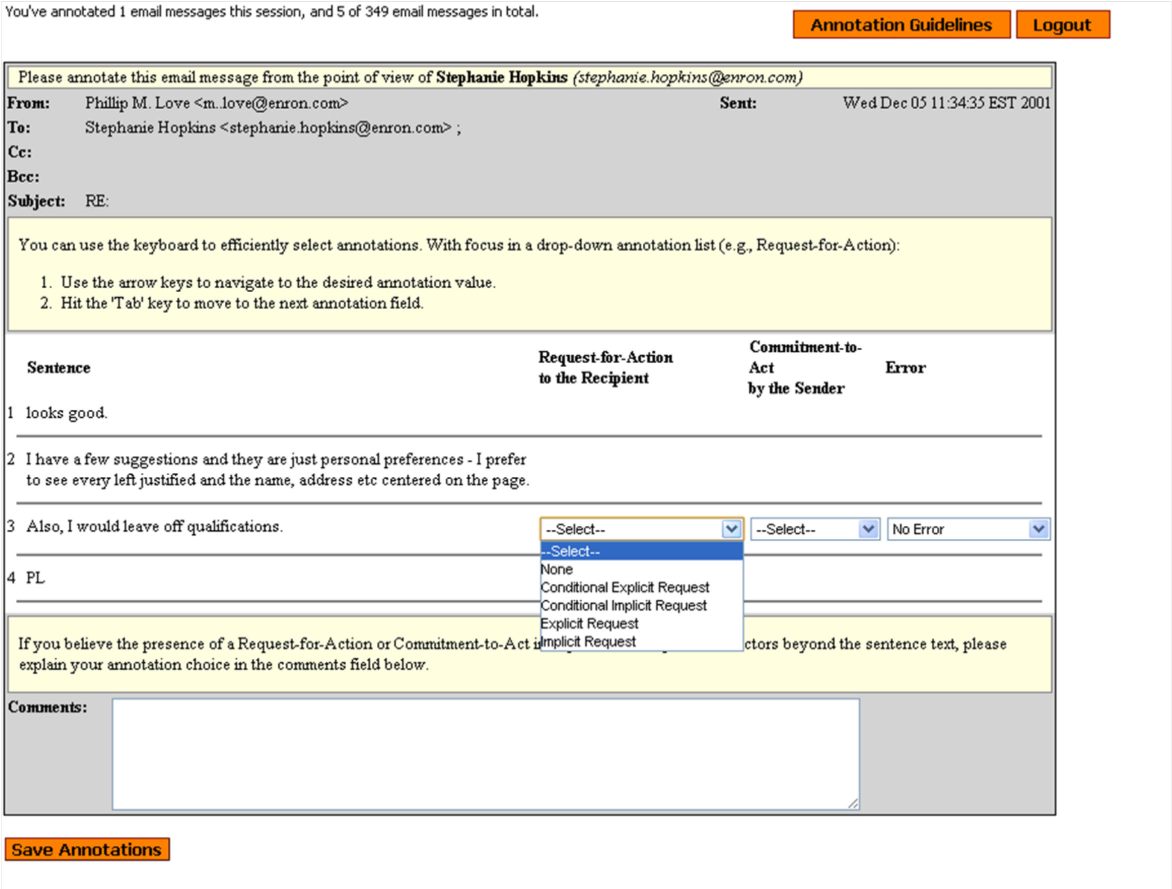
For practical reasons in regards to access to appropriate data, none of our annotation tasks included any information about the relationships between senders and recipients. We do, however, return to this issue in our discussion of in-context annotation in Section 6.3.2.

Analysis of disagreements from our first experiment led us to include greater detail and guidance for classifying what we began to think of as edge cases. Our revised guidelines for the second annotation task, which are included in Section B.2.1, specified that both requests and commitments should be explicitly classified as conditional or unconditional, to emphasise the inclusion of conditional requests and commitments. We also asked annotators to mark requests and commitments as either implicit or explicit, aiming to encourage more careful consideration of implicit speech acts. These two changes were made to address the major classes of disagreement in our first annotation task. We consider that our first annotation task results demonstrate that annotators require guidance about how to classify conditional requests and commitments to achieve even moderate inter-annotator agreement. Others have since replicated this finding (Scerri et al., 2008).

The second annotation task (*Sentence 2* in Table 3.1) is described in Section B.2, and was performed using a modified version of the same web-based annotation tool. The modifications to the tool were in the categories with which each sentence was marked, as shown in Figure 3.16. Rather than choosing the relative strength of each request and commitment marked, annotators were required to identify both the conditionality (CONDITIONAL or UNCONDITIONAL) and explicitness (IMPLICIT or EXPLICIT) of each request and commitment marked. Conditionality is as discussed in Section 3.1.7. Explicit or direct requests were defined as those that state the request explicitly, either in imperative or interrogative form, while implicit or indirect requests do not. Similarly, explicit or direct commitments state the promise for future action explicitly. Implicit or indirect commitments do not state the promise to complete future action. We elaborate on these definitions in Section B.2.1.

Another significant difference was that we did not allow all sentences from each email message to be annotated. Given the low density of requests and commitments observed on our first annotation task, the intention was to increase the number of positive requests and commitments identified per unit of annotation. Although all author text content was shown to provide context for interpretation to each annotator, only a subset of sentences were selected to be annotated. Sentences were selected for annotation based on cue phrases correlated with requests and commitments. These cue phrases were derived from the data annotated in our first annotation experiment. In total, annotators marked 750 sentences drawn from 350 email messages which were themselves randomly sampled from the Enron corpus.

While acknowledging that the results are not directly comparable, our inter-annotator



You've annotated 1 email messages this session, and 5 of 349 email messages in total.

Please annotate this email message from the point of view of **Stephanie Hopkins** *(stephanie.hopkins@enron.com)*

From:	Phillip M. Love <m..love@enron.com>	**Sent:**	Wed Dec 05 11:34:35 EST 2001
To:	Stephanie Hopkins <stephanie.hopkins@enron.com> ;		
Cc:			
Bcc:			
Subject:	RE:		

You can use the keyboard to efficiently select annotations. With focus in a drop-down annotation list (e.g., Request-for-Action):

1. Use the arrow keys to navigate to the desired annotation value.
2. Hit the 'Tab' key to move to the next annotation field.

	Sentence	**Request-for-Action to the Recipient**	**Commitment-to-Act by the Sender**	**Error**
1	looks good.			
2	I have a few suggestions and they are just personal preferences - I prefer to see every left justified and the name, address etc centered on the page.			
3	Also, I would leave off qualifications.	--Select--	--Select--	No Error
4	PL			

If you believe the presence of a Request-for-Action or Commitment-to-Act i ctors beyond the sentence text, please explain your annotation choice in the comments field below.

Comments:

Figure 3.16: Our web-based annotation tool used in our sentence-level annotation tasks.

agreement scores increased for both requests and commitments compared with our first experiment, particularly for commitments. The improved agreement seemed to benefit from the better guidance about how to annotate conditional commitments (and to a lesser degree, requests).

Binary three-way κ agreement scores for requests increased from 0.78 to 0.80, and for commitments from 0.54 to 0.74. At the fine-grained level, requiring consensus on conditionality and explicitness, three-way request agreement remained the same as fine-grained agreement for strength in our first annotation experiment at 0.60, while fine-grained commitment agreement increased significantly from 0.37 to 0.70. Of course, for the fine-grained κ scores, we are not comparing like with like, since the original scores are based on agreements about strength, and the current scores on agreement about explicitness and conditionality.

The biggest improvements are in the agreement for identifying commitments, both at the binary level (*Does this sentence contain any type of commitment?*) and at the finer-grained level (*Does this sentence contain a conditional/unconditional implicit/explicit commitment?*).

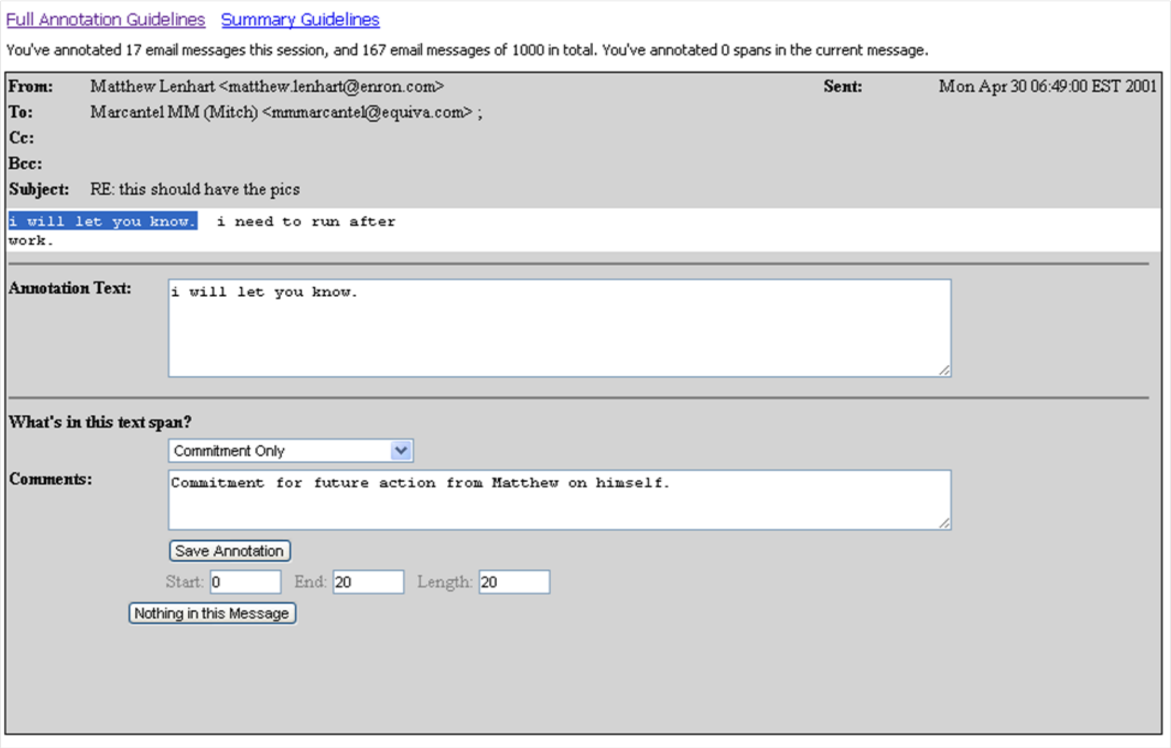
	3-way Binary κ		*Fine-grained κ*	
	All	Non-Error	All	Non-Error
Request	0.80	0.84	0.60	0.62
Commitment	0.74	0.76	0.70	0.72

Table 3.2: Three-way κ agreements for classifying requests and commitments in the second annotation experiment, contrasting results where sentences marked with errors are included and excluded.

In our second annotation task, many cases of disagreement stemmed from the way annotators marked quoted text and forwarded material in email messages. Annotator B generally marked requests and commitments in quoted and forwarded text as if they were written for the specified recipient. Annotator C generally did not mark requests or commitments in these text sections, and Annotator A varied across different cases.

We attempted to remove forwarded and quoted reply content from messages before they were presented to annotators. Where such text remained, annotators were instructed to indicate failed pre-processing using an appropriate error category (e.g., *Forwarded Content*). Given the significant disagreement observed in interpreting obligation acts in forwarded and quoted content, we calculated agreement scores after excluding all text marked by at least one annotator with an error. Table 3.2 shows fine-grained and binary three-way κ agreement scores after sentences marked with an error have been removed. These results, which are better for binary and fine-grained agreement across both requests and commitments, suggest that agreement on sentences of interest is actually higher than our original results suggest. A non-trivial fraction of disagreements are related to processing errors in our pre-annotation pipeline that attempts to remove non-author text and segment email text into sentences. These errors are at least partially attributable to our annotation process, rather than necessarily being genuine disagreements about interpreting requests and commitments.

Our third, fourth and fifth annotation tasks (*Message 1*, *Message 2* and *Message 3* in Table 3.1) continued from the approach in our first two annotation tasks, with email annotated at the message level. The *Message 1* annotation task recorded three-way agreement of $\kappa = 0.84$ for requests and $\kappa = 0.78$ for commitments. The second message experiment (*Message 2* in Table 3.1) introduced two non-expert annotators, and saw maximum three-way agreement of $\kappa = 0.79$ for requests and $\kappa = 0.80$ for commitments. Five-way agreement, however, was markedly lower, with $\kappa = 0.66$ for both requests and commitments. Much of this disagreement stemmed from confusion between phatic and obligation acts, leading us to more carefully define the distinction in the third message-level experiment (*Message 3* in Table 3.1). This saw three-way agreement lift back to $\kappa = 0.85$ for requests and $\kappa = 0.86$ for commitments. The annotated data from our *Message 3* experiment was then used as training data to build our message-level classifiers that we present in Chapter 4.



You've annotated 17 email messages this session, and 167 email messages of 1000 in total. You've annotated 0 spans in the current message.

From: Matthew Lenhart <matthew.lenhart@enron.com> **Sent:** Mon Apr 30 06:49:00 EST 2001
To: Marcantel MM (Mitch) <mmmarcantel@equiva.com> ;
Cc:
Bcc:
Subject: RE: this should have the pics

```
i will let you know.  i need to run after
work.
```

Annotation Text: i will let you know.

What's in this text span? Commitment Only

Comments: Commitment for future action from Matthew on himself.

Start: 0 End: 20 Length: 20

Figure 3.17: The web-based annotation tool we developed and used in our span annotation tasks.

3.2.4 Our Free Text Span Annotation Tasks

As requests and commitments are not realised with fixed units of text, we also completed two free text annotation tasks that required annotators to both identify the extent of text they wished to annotate, and then choose the appropriate label. We refer to these tasks as span annotation tasks. They avoid problems with segmenting email messages, which frequently contain sentence fragments and other non-sentential text, into sentences, by allowing annotators to mark text spans that do not necessarily correlate with paragraph or sentence boundaries.

Both our span annotation tasks were performed using another custom-built annotation tool, shown in Figure 3.17. This tool allows annotators to mark textual units of any size, from a single character up to the entire content of an email message, and then to label the span as a request, a commitment, a phatic obligation act, or combinations of these. As shown in Figure 3.17, an annotator first highlights a span of text in the message body or subject, and then selects a label from the closed set of available annotations.

Two independent annotators used our span annotation tool to mark all request and commitment spans in a collection of 1000 business email messages, drawn from the Enron email corpus. The annotation guidelines which they were given are provided

				2-way F-score	
Experiment	Annotators	Msgs	Spans Marked	Request	Commit
Span 1	3	50	221	0.91	0.86
Span 2	2	1000	3397	0.86	0.72

Table 3.3: F-score agreement from manual free text span email annotation experiments.

in Appendix A.1. The annotated spans were used to create gold-standard training and evaluation corpora for both our paragraph-level and sentence-level request and commitment classifiers. We describe how this is done, including how we deal with annotation disagreements, later in this section.

The resulting annotated corpus contains 3397 labelled spans of text. The longest annotated span was 2198 characters long (379 words across 23 sentences), containing a sequence of instructions to follow in the event of receiving a package suspected of containing Anthrax, marked as a request; the shortest agreed span was one word of 7 characters, consisting of the utterance *cancel?*, in response to an interview request email, which was marked as a request by both annotators.

Because annotators must both identify the appropriate span of text, and then apply a label to the span, calculating inter-annotator agreement involves two distinct aspects of agreement that must each be measured. We use F-score to measure inter-annotator agreement, following the standard approach used in named entity recognition (NER) research, which also deals with marking and labelling free spans of text (Hachey, Alex, and Becker, 2005). This is in contrast, to our message- and sentence-level annotation tasks where the units of annotation were fixed. Once we have identified text spans that are considered the same, we then measure agreement using a pairwise F-score metric. With no obvious gold standard available, we again follow the standard approach used in NER research by treating the annotations of one annotator as the 'correct' annotations, and measuring agreement from the second annotator against those of our first annotator. Note that it does not matter which annotator's annotations we choose to be the 'gold standard', since reversing the choice simply reverses our values for precision and recall, which results in the same overall F-score. Where more than two annotators are involved, as in our *Span 1* experiment, we calculate each pairwise F-score, and then average the results.

As noted above, to calculate agreement, we need to first establish when two annotators are marking the 'same' span of text, and then to establish agreement in labelling those agreed spans. A reasonable question is how we measure text spans and labels which are the 'same'. The naïve answer to both these questions is to consider only exact matches as agreement, but, as we discuss in Section 5.1, this approach is overly restrictive and does not match human intuitions about agreement. To measure whether two spans of text are the same, we use a window of overlapping words with a threshold of overlapping words to decide when two spans are equivalent. To measure whether two labels are the same, we consider agreement only for the obligation act of interest. This means, for example, that if we are looking for agreed requests, and annotator A has marked a span with a request, and annotator B with a request and a commitment, then we would consider the span to have an agreed request label. We provide more detail on how these metrics are calculated in Section 5.1 where we present our work on automated, fine-grained request and commitment classification, for which the annotated corpus generated in the *Span 2* experiment forms our training dataset.

Our inter-annotator agreement F-scores for our two span annotation experiments are shown in Table 3.3.

3.3 Functional Features of Requests and Commitments

As can be readily discerned from the annotation experiments and annotation agreement described in Section 3.2, one of the key findings from our empirical examination of requests and commitments is that the utterances which convey these actionable intentions are frequently open to differing interpretation. While this should come as no surprise to anyone familiar with linguistic work in the field of pragmatics, it stands in stark contrast to the simplicity assumed by almost all previous work that we have surveyed which attempts to automate the detection of task-related speech acts.

In this section, we provide detailed examples and guidance for a range of features and phenomena in request and commitment utterances that we uncovered during our empirical analysis. The phenomena that we cover are:

- Phatic obligation acts that look like obligation acts, but do not convey significant obligation and are primarily intended to make an email message appear professional, polite, or to conform with social norms. We cover phatic acts in Section 3.3.1.

- Obligation acts that require ongoing action. These acts have different characteristics than acts that request one-time action. Ongoing actions, for example, often cannot meaningfully be added to a task list and completed. We discuss the distinction between these types of obligation acts in Section 3.3.2.

- Conversations where the creation and activation of the obligation is separated over different messages. Examples include offers, invitations and threats, where the obligation is created by these acts, but triggered by action or response from the recipient at some later stage. We discuss these acts in Section 3.3.3.

- Transitive obligation acts that are uttered by a third party and conveyed by an email sender. Reported requests and commitments and forwarded obligation acts are common examples that we discuss when we examine this phenomenon in Section 3.3.4.

- Processes and instructions that can create ambiguity around whether their content is actionable or to be filed for later use. We discuss these types of utterances in Section 3.3.5.
- Obligations for inaction require recipients *not* to act. Such obligation acts do not fit neatly into a user's task list, and can cause confusion for annotators. We discuss these acts in Section 3.3.6.
- Obligation acts that are frequently supported by additional utterances providing rhetorical support, such as justification or elaboration for a request or a commitment. In Section 3.3.7, we discuss how to deal with supporting text when identifying requests and commitment.
- The genre and functional layout of content in an email message. These can have a significant effect on how specific parts of its content are interpreted from the perspective of identifying requests and commitments. We discuss these in Section 3.3.8.
- Email messages that have documents and other media attached to them. Should requests that make reference to reading, saving or otherwise acting on these attached documents be marked as obligation acts? We consider these issues in Section 3.3.9.

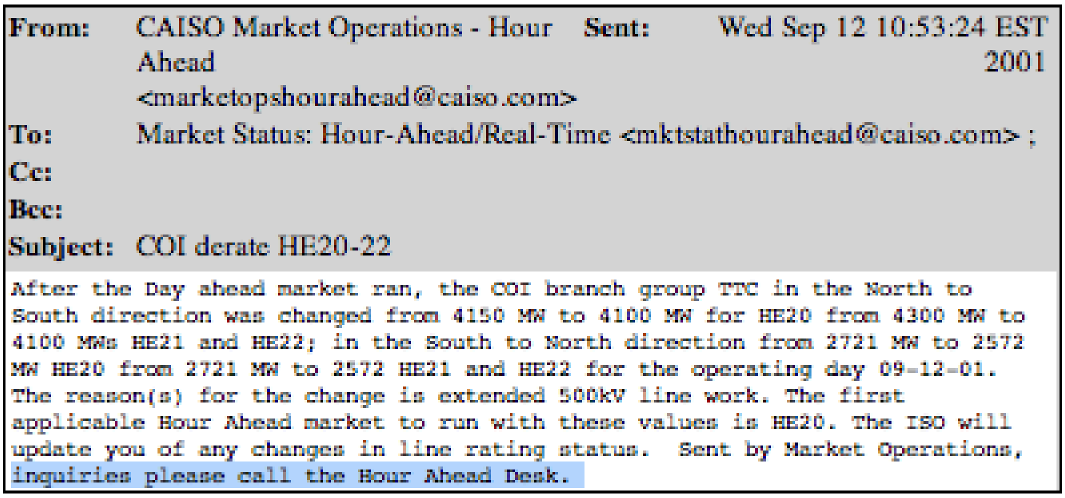
Together, these phenomena blur the boundaries between utterances that represent obligation acts and those that do not, and reveal the complexity that underlies obligation acts in real-world usage. An important implication of the phenomena and features identified in this section is that they should be explicitly considered before one attempts to either manually or automatically identify requests or commitments in email text.

3.3.1 Phatic Obligation Acts

Utterances in the style of *Let me know if you have any questions* are very common in workplace email. In some cases, these utterances actually carry the illocutionary force of a request for the recipient to act and a commitment for the author to then respond. In many cases, however, their presence appears intended to make the message appear professional, polite or to conform with social norms. In such cases these utterances are a form of PHATIC communication (Malinowski, 1923).

Phatic language has meaning decoupled from its literal meaning, and in this way presents as a phenomenon with much in common with indirect speech acts. The purpose of phatic language is often to indicate professionalism, friendliness or politeness (Laver, 1981). When a sender asks *How are you?* in an email message, they are often asking this phatically, as an extended form of greeting rather than as an interrogative request for information. The sender often expects no response to that specific utterance.

In the context of identifying requests and commitments, if an utterance resembles a request or a commitment, either through its surface form or by convention, but does not



From: CAISO Market Operations - Hour Ahead <marketopshourahead@caiso.com> **Sent:** Wed Sep 12 10:53:24 EST 2001
To: Market Status: Hour-Ahead/Real-Time <mktstathourahead@caiso.com> ;
Cc:
Bcc:
Subject: COI derate HE20-22

```
After the Day ahead market ran, the COI branch group TTC in the North to
South direction was changed from 4150 MW to 4100 MW for HE20 from 4300 MW to
4100 MWs HE21 and HE22; in the South to North direction from 2721 MW to 2572
MW HE20 from 2721 MW to 2572 HE21 and HE22 for the operating day 09-12-01.
The reason(s) for the change is extended 500kV line work. The first
applicable Hour Ahead market to run with these values is HE20. The ISO will
update you of any changes in line rating status.  Sent by Market Operations,
inquiries please call the Hour Ahead Desk.
```

Figure 3.18: An email containing a unanimously agreed phatic request, as highlighted.

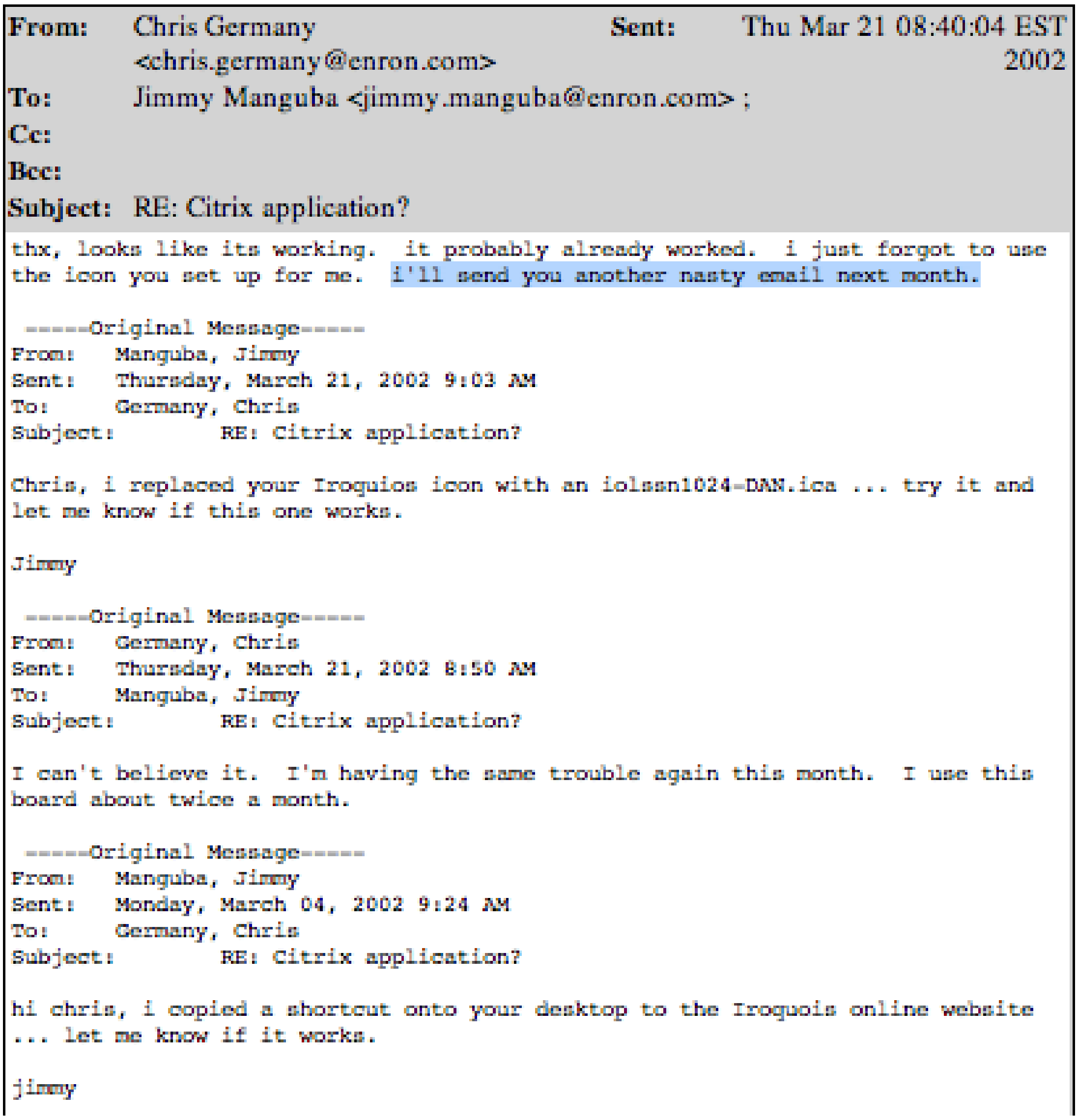
convey any significant obligation, then we label the utterance as a phatic request or a phatic commitment.[13] Corston-Oliver et al. (2004) observed similar utterances in their own corpus of workplace email, labelling them formulaic endings. Such utterances can, however, appear anywhere in an email message; we therefore prefer the more general phatic obligation act label.

An example phatic request is highlighted in Figure 3.18. In this message, the utterance *inquiries please call the Hour Ahead Desk* was considered to be a phatic request by all our annotators. Despite resembling a request, its form is very formulaic, with identical utterances appearing in multiple email messages. More importantly, the utterance does not appear to actually place any obligation on the recipient to act; while clearly open to the possibility of the recipient replying, it is unlikely that the sender of the message actually expects a reply.

For completeness, Fig 3.19 shows a phatic commitment. While the highlighted utterance *i'll send you another nasty email next month* shares the form of a commitment, it is clearly meant as a joke, and thus carries no actual commissive intent.

Like Corston-Oliver et al. (2004), we believe that distinguishing between phatic obligation acts and actual obligation acts requires consideration of the context of the entire message. In our annotation experiments, annotators were instructed to exercise their own judgement to distinguish when utterances such as the following function as phatic utterances, and when they are expected to elicit action or response from the

[13] In earlier work (e.g., (Lampert, Dale, and Paris, 2008a; Lampert, Dale, and Paris, 2010)), we referred to this class of utterances as Pleasantries. We found the term Pleasantry caused confusion for some annotators, given its association with non-directive utterances such as *Thanks for your message*. We avoid this confusion by using the terms phatic request and phatic commitment in its place.



From: Chris Germany <chris.germany@enron.com> **Sent:** Thu Mar 21 08:40:04 EST 2002
To: Jimmy Manguba <jimmy.manguba@enron.com> ;
Cc:
Bcc:
Subject: RE: Citrix application?

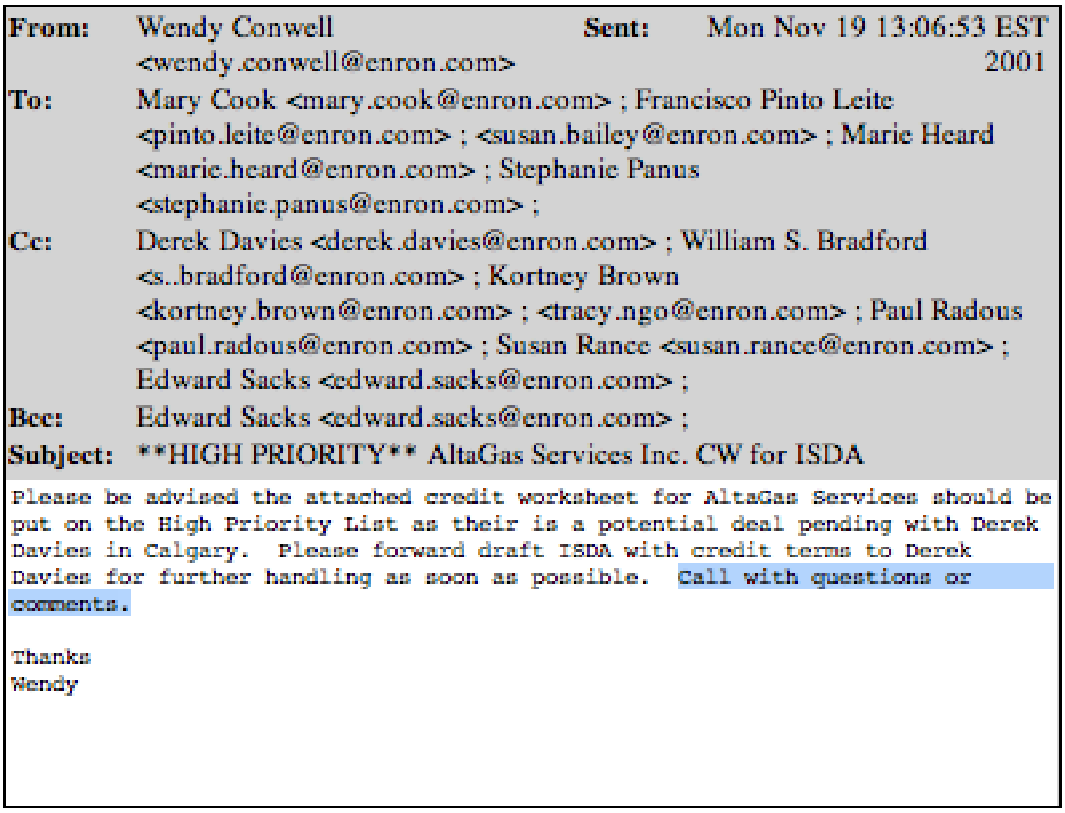
```
thx, looks like its working.  it probably already worked.  i just forgot to use
the icon you set up for me.  i'll send you another nasty email next month.

 -----Original Message-----
From:   Manguba, Jimmy
Sent:   Thursday, March 21, 2002 9:03 AM
To:     Germany, Chris
Subject:        RE: Citrix application?

Chris, i replaced your Iroquios icon with an iolssn1024-DAN.ica ... try it and
let me know if this one works.

Jimmy

 -----Original Message-----
From:   Germany, Chris
Sent:   Thursday, March 21, 2002 8:50 AM
To:     Manguba, Jimmy
Subject:        RE: Citrix application?

I can't believe it.  I'm having the same trouble again this month.  I use this
board about twice a month.

 -----Original Message-----
From:   Manguba, Jimmy
Sent:   Monday, March 04, 2002 9:24 AM
To:     Germany, Chris
Subject:        RE: Citrix application?

hi chris, i copied a shortcut onto your desktop to the Iroquois online website
... let me know if it works.

jimmy
```

Figure 3.19: An email containing a unanimously agreed phatic commitment.

recipient or sender:

(3.77) *Let me know if you have any questions*

Based on this approach, we found that considerable disagreement between our annotators stemmed from confusion between phatic requests and actual requests. We experimented with creating separate categories for phatic requests and phatic commitments to help focus annotators on distinguishing obligation acts from phatic acts, and to help us quantify the disagreement that arises from this phenomenon. In our



From: Wendy Conwell <wendy.conwell@enron.com> **Sent:** Mon Nov 19 13:06:53 EST 2001
To: Mary Cook <mary.cook@enron.com> ; Francisco Pinto Leite <pinto.leite@enron.com> ; <susan.bailey@enron.com> ; Marie Heard <marie.heard@enron.com> ; Stephanie Panus <stephanie.panus@enron.com> ;
Cc: Derek Davies <derek.davies@enron.com> ; William S. Bradford <s..bradford@enron.com> ; Kortney Brown <kortney.brown@enron.com> ; <tracy.ngo@enron.com> ; Paul Radous <paul.radous@enron.com> ; Susan Rance <susan.rance@enron.com> ; Edward Sacks <edward.sacks@enron.com> ;
Bcc: Edward Sacks <edward.sacks@enron.com> ;
Subject: **HIGH PRIORITY** AltaGas Services Inc. CW for ISDA

```
Please be advised the attached credit worksheet for AltaGas Services should be
put on the High Priority List as their is a potential deal pending with Derek
Davies in Calgary.  Please forward draft ISDA with credit terms to Derek
Davies for further handling as soon as possible.  Call with questions or
comments.

Thanks
Wendy
```

Figure 3.20: An email containing a more ambiguous utterance that may be a phatic request: *Call with questions or comments.*

Span 2 experiment, disagreements between phatic and actual obligation acts impact our overall F-score by approximately 1.5 points.

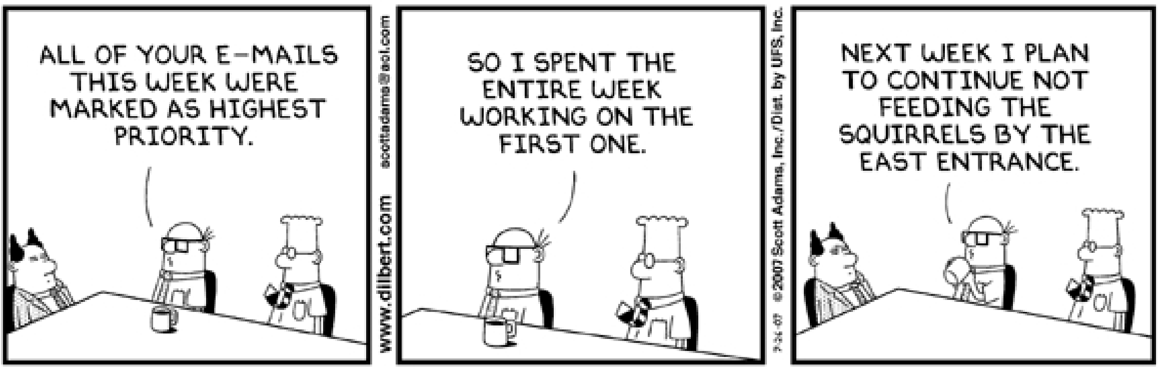
Even with explicit guidance to consider the entire context of a message in interpreting utterances, there remain cases of unresolved disagreement between our annotators around the interpretation of specific utterances. An example is the following utterance from the email in Figure 3.20:

(3.78) *Call with questions or comments*

The directness of this utterance led to differing interpretations by our annotators of whether it represents a genuine request that is attempting to elicit a response.

Taking all the above into account, our guidance for interpreting with phatic obligation acts is that:

Phatic acts are not obligation acts.


ALL OF YOUR E-MAILS THIS WEEK WERE MARKED AS HIGHEST PRIORITY.
www.dilbert.com scottadams@aol.com
SO I SPENT THE ENTIRE WEEK WORKING ON THE FIRST ONE.
7-24-07 © 2007 Scott Adams, Inc./Dist. by UFS, Inc.
NEXT WEEK I PLAN TO CONTINUE NOT FEEDING THE SQUIRRELS BY THE EAST ENTRANCE.

© Scott Adams, Inc./Dist. by UFS, Inc.

Figure 3.21: An example of an ongoing request for inaction in email.

3.3.2 Ongoing Action and One-time Action

In previous work analysing speech acts in email, request acts have sometimes been defined as utterances that require a recipient to add an action to their task list (e.g., (Corston-Oliver et al., 2004; Faulring et al., 2010)). Such a definition, however, covers only certain classes of requests. Figure 3.21 shows a humorous example of a request that cannot be simply added to a task list and completed.

Importantly, there are a wide range of requests that do not satisfy an informal ‘to-do list test’. Consider, for example, the following request:

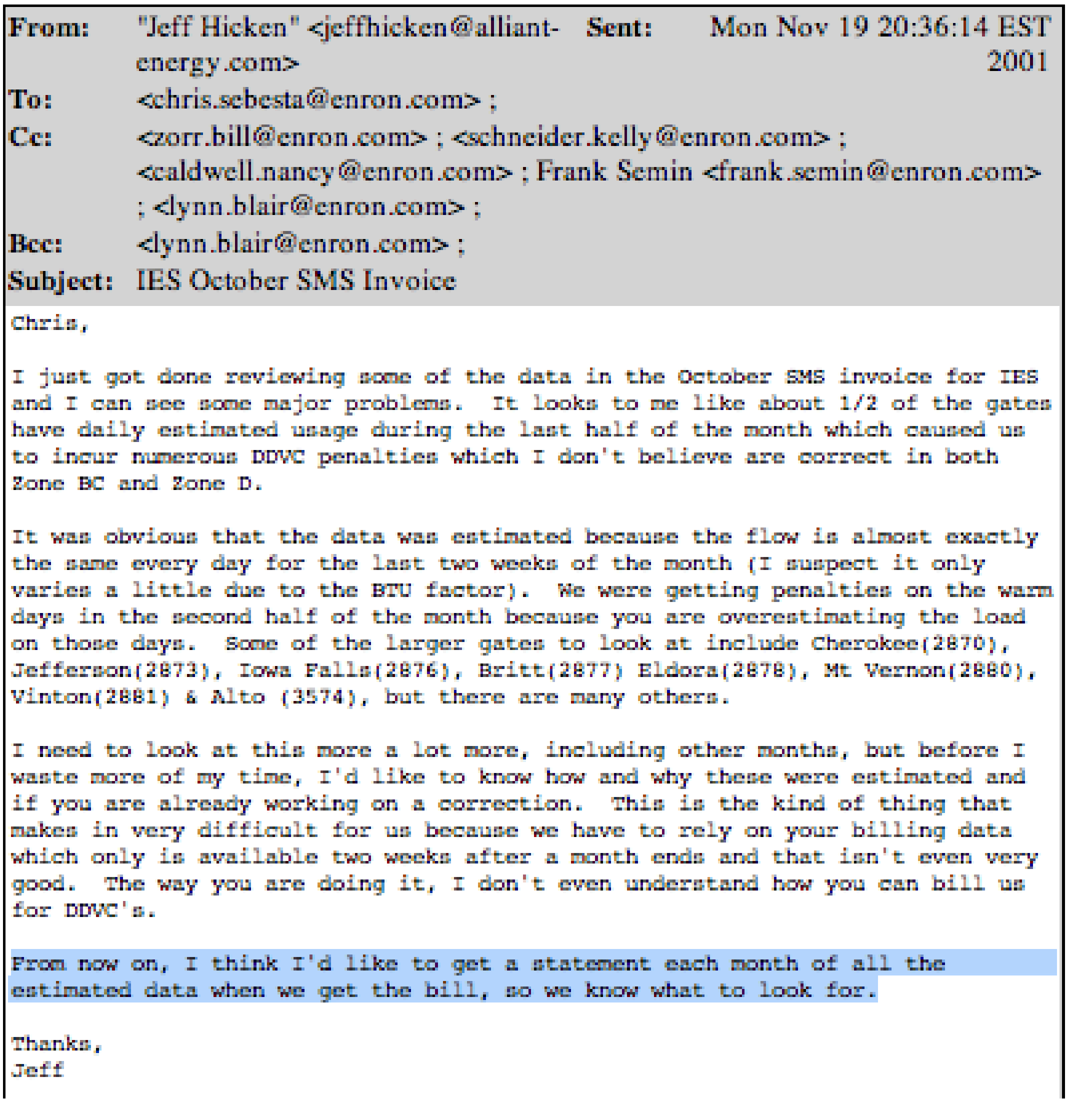
(3.79) *Make sure you always send a copy to Wendy too.*

The request in Example (3.79) requests *ongoing* action, in contrast with requests that require a *one-time* action. A one-time action can be identified and undertaken, after which the original request is considered to have been satisfied; an ongoing action, however, requires continual action, and often has a less definite sense of completion.

A one-time version of the request from Example (3.79) is shown in Example (3.80). This form of the request, while differing only by one word, is completable, while the request in Example (3.79) is something that carries an ongoing obligation.

(3.80) *Make sure you send a copy to Wendy too.*

In this section, we focus on the distinction between obligation acts that impose an obligation for ongoing action or inaction, which we call ONGOING REQUESTS or ONGOING COMMITMENTS (e.g., a request for behavioural change), and requests or commitments that impose an obligation for once-off action that we call ONE-TIME REQUESTS or ONE-TIME COMMITMENTS. Ongoing obligation acts may or may not be unbound in time; some apply for an indefinite or undefined period of time, as in Example (3.79), while others have a defined period of action, as in Example (3.81).



From: "Jeff Hicken" <jeffhicken@alliant-energy.com> **Sent:** Mon Nov 19 20:36:14 EST 2001
To: <chris.sebesta@enron.com> ;
Cc: <zorr.bill@enron.com> ; <schneider.kelly@enron.com> ; <caldwell.nancy@enron.com> ; Frank Semin <frank.semin@enron.com> ; <lynn.blair@enron.com> ;
Bcc: <lynn.blair@enron.com> ;
Subject: IES October SMS Invoice

Chris,

I just got done reviewing some of the data in the October SMS invoice for IES and I can see some major problems. It looks to me like about 1/2 of the gates have daily estimated usage during the last half of the month which caused us to incur numerous DDVC penalties which I don't believe are correct in both Zone BC and Zone D.

It was obvious that the data was estimated because the flow is almost exactly the same every day for the last two weeks of the month (I suspect it only varies a little due to the BTU factor). We were getting penalties on the warm days in the second half of the month because you are overestimating the load on those days. Some of the larger gates to look at include Cherokee(2870), Jefferson(2873), Iowa Falls(2876), Britt(2877) Eldora(2878), Mt Vernon(2880), Vinton(2881) & Alto (3574), but there are many others.

I need to look at this more a lot more, including other months, but before I waste more of my time, I'd like to know how and why these were estimated and if you are already working on a correction. This is the kind of thing that makes in very difficult for us because we have to rely on your billing data which only is available two weeks after a month ends and that isn't even very good. The way you are doing it, I don't even understand how you can bill us for DDVC's.

From now on, I think I'd like to get a statement each month of all the estimated data when we get the bill, so we know what to look for.

Thanks,
Jeff

Figure 3.22: An email containing an ongoing request.

(3.81) *While Liz is on leave, make sure you always send a copy to Wendy too.*

Instructions, process descriptions and requests for inaction are common sources of ongoing requests. Figure 3.30, for example, contains the utterance *Please, don't distribute it*, which we interpret as a REQUEST FOR INACTION. Depending on the context, this particular request may be interpreted as open-ended in time, or may apply only during a fixed time period, such as prior to a legal case being settled. In either case, we consider this an ongoing request. Another example of an ongoing request is shown below in Figure 3.22. The ongoing request is found at the end of the message: *From now on, I think I'd like to get a statement each month of all the estimated data when we get the* *bill, so we know what to look for.* Its ongoing nature is flagged through the use of the cue phrase *From now on.*

Vine (2004) explores a related distinction in her analysis of spoken requests in the workplace. She distinguishes between NOW requests, which require immediate compliance, and LATER requests, where the completion of action is delayed to another place and time. This distinction captures a slightly different aspect of variance, and is also less overt in email communication where, due to its asynchronous nature, the sender and recipient rarely write and read the message at the same time. We could quite readily distinguish NOW ONE-TIME REQUESTS, such as *Please send me the document ASAP*, which require action now, from LATER ONE-TIME REQUESTS, such as *Please send me the document on Friday*, which require action at some future time. The same distinction could be drawn for ongoing requests, as shown in Examples (3.82) and (3.83).

(3.82) *Always include the office manager when you distribute this document.*

(3.83) *From next month, you'll also need to include the office manager when you distribute this document.*

Exactly the same distinction between ongoing and one-time future actions can also be drawn for commitments. An example of an ongoing commitment is: *I promise I won't speak out of turn in future meetings.* Such a commitment cannot be completed and considered fulfilled; it is an ongoing and open-ended commitment to behavioural change. In contrast, a similar one-time commitment is: *I promise I won't speak out of turn in our next meeting.*

Distinguishing between ongoing and one-time actions has important implications at the application level for identifying one-time requests and commitments that might be migrated to a to-do list. ongoing requests and commitments, however, might be highlighted *in situ* within the text of a message to draw attention to them as actionable content, but would not usually be imported into task lists, calendars or other structured task management views.

After considering all of the above, our guidance for interpreting obligation acts requiring once-off and ongoing action is that:

> **One-time and ongoing requests and commitments are all obligation acts that we wish to identify.**

3.3.3 Separate Creation and Activation of Obligation

Consider the following sequence of utterances each extracted from a separate email message within the same conversation thread:

(3.84) Ronan: *Would you like me to help with the report?*

(3.85) Finn: *Sounds great!*

It is interesting to ask whether the utterance in Example (3.85) acts as a request for Ronan. It seems unambiguous that Ronan is obligated to help with the report after Finn includes this utterance in his response. The obligation for Ronan to help with the report, however, is not defined or created by the utterance in Example (3.85). Instead, the obligation is created when Ronan makes his preceding offer, shown in Example (3.84). The commissive obligation created by Ronan's offer is then triggered by Finn's acceptance of Ronan's offer.

In many requests and commitments, there is no distinction between defining and activating or triggering the associated obligation, as both are realised within a single utterance. In such obligation acts, the obligation is created as soon as the utterance is transmitted. For example, as soon as someone asks *What time does the meeting start on Friday?* the obligation to respond is placed on the recipient or hearer. As we have just seen above however, there are some utterances where the creation or definition of an obligation and the trigger for activating that obligation are achieved in separate utterances, often within separate email messages.

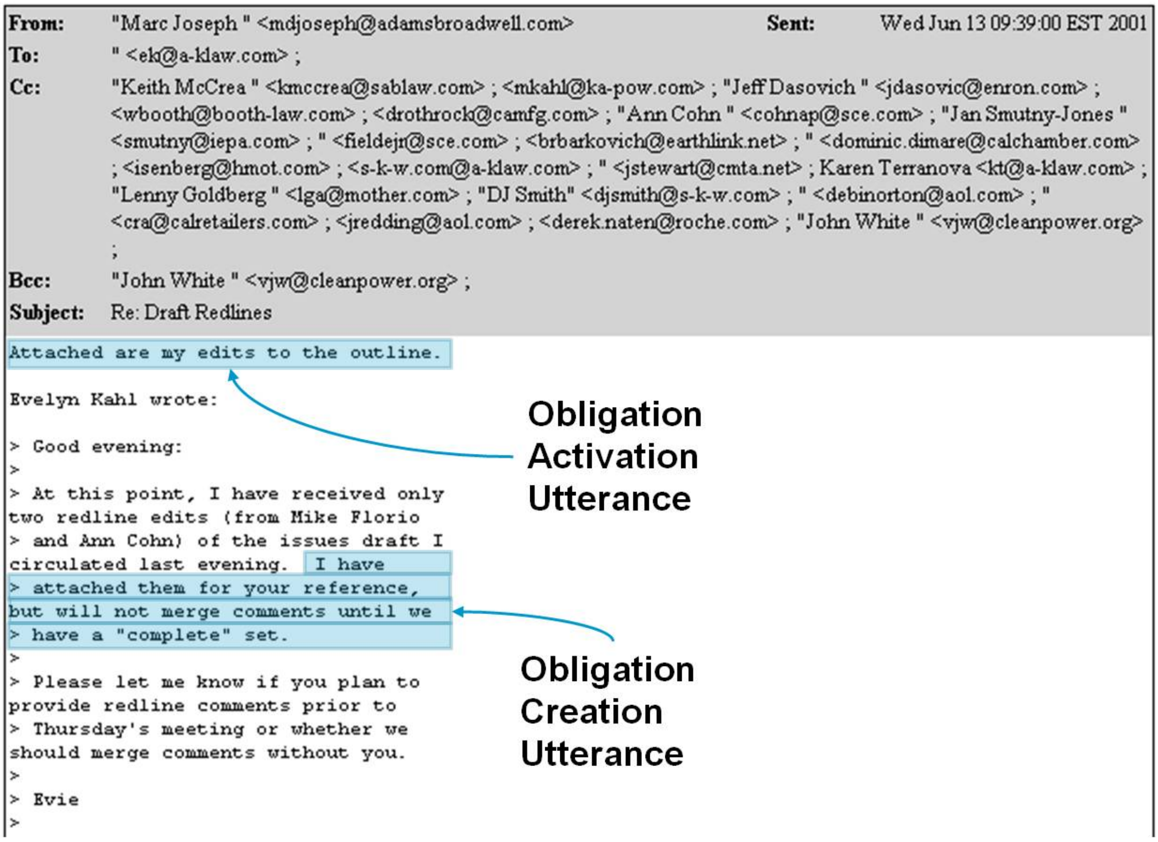
Figure 3.23, provides a second example of this phenomenon. We see a commitment in the quoted message, an earlier message from Evelyn to a set of recipients including Marc Joseph, that creates an obligation for Evelyn. The specific commitment utterance is:

(3.86) *I have attached them for your reference, but will not merge comments until we have a "complete" set.*

The obligation associated with the commitment in Example (3.86) is not activated until people actually contribute comments. The response in the current message from Marc Joseph provides comments that then activate or trigger the obligation associated with the earlier commitment from Evelyn. Note that the utterance or action of providing comments does not itself *create* an obligation for Evelyn; it is not functioning as a request. Rather, when Marc provides his comments, his message activates the previously created obligation in Evelyn's own commitment in her earlier message.

Analysis of our annotated datasets suggests that utterances which exhibit this temporal separation of defining and imposing the obligation frequently function simultaneously as both a request and a conditional commitment. Offering and inviting, along with some other more specialised speech acts such as tendering or bidding, can all exhibit this property of simultaneously acting as both a request and a commitment, and are sometimes said to act as hybrid speech acts, in that they combine commissive (commitment) and directive (request) force. Some linguists have moved to identify new categories of speech acts (e.g., commissive directives from (Hancher, 1979)) to describe these hybrid acts, though none of these categories have been widely adopted in the speech act community.

In Sections 3.3.3.1 and 3.3.3.2 we discuss offers and invitations in more detail, as these are the most frequently occurring acts in workplace email which exhibit both simultaneous request and commitment force, and the separate creation and activation



From: "Marc Joseph " <mdjoseph@adamsbroadwell.com> Sent: Wed Jun 13 09:39:00 EST 2001
To: " <ek@a-klaw.com> ;
Cc: "Keith McCrea " <kmccrea@sablaw.com> ; <mkahl@ka-pow.com> ; "Jeff Dasovich " <jdasovic@enron.com> ; <wbooth@booth-law.com> ; <drothrock@camfg.com> ; "Ann Cohn " <cohnap@sce.com> ; "Jan Smutny-Jones " <smutny@iepa.com> ; " <fieldejr@sce.com> ; <brbarkovich@earthlink.net> ; " <dominic.dimare@calchamber.com> ; <isenberg@hmot.com> ; <s-k-w.com@a-klaw.com> ; " <jstewart@cmta.net> ; Karen Terranova <kt@a-klaw.com> ; "Lenny Goldberg " <lga@mother.com> ; "DJ Smith" <djsmith@s-k-w.com> ; " <debinorton@aol.com> ; " <cra@calretailers.com> ; <jredding@aol.com> ; <derek.naten@roche.com> ; "John White " <vjw@cleanpower.org> ;
Bcc: "John White " <vjw@cleanpower.org> ;
Subject: Re: Draft Redlines

```
Attached are my edits to the outline.

Evelyn Kahl wrote:

> Good evening:
>
> At this point, I have received only
two redline edits (from Mike Florio
> and Ann Cohn) of the issues draft I
circulated last evening.  I have
> attached them for your reference,
but will not merge comments until we
> have a "complete" set.
>
> Please let me know if you plan to
provide redline comments prior to
> Thursday's meeting or whether we
should merge comments without you.
>
> Evie
>
```

Obligation Activation Utterance
Obligation Creation Utterance

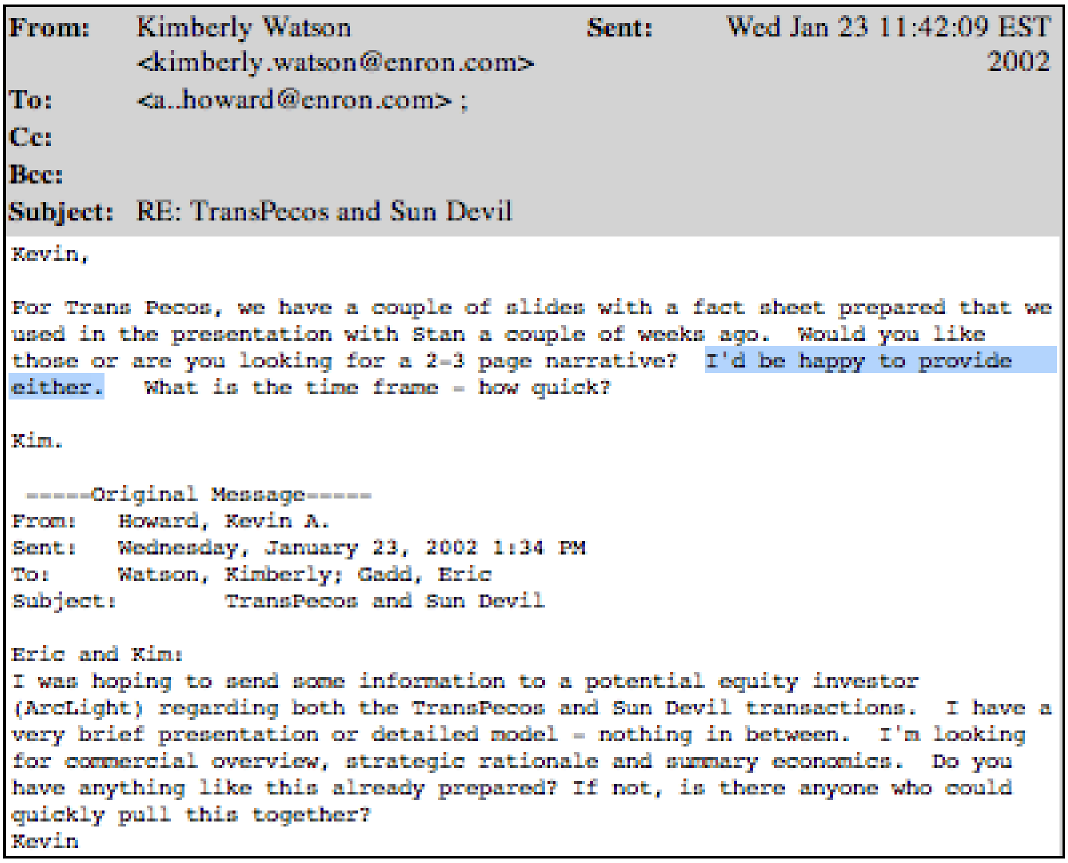
Figure 3.23: Illustrating the phenomenon of separation of obligation creation and activation.

of obligation that we have presented here. We also briefly discuss threats in Section 3.3.3.3, in order to distinguish them from offers, with which they have much in common.

3.3.3.1 Offers

Consider the highlighted utterance shown in Figure 3.24. We refer to this utterance, *I'd be happy to provide either*, as an offer. Offers are interesting speech acts, as they function as both a conditional commitment and a conditional request. The conditional request stems from the recipient of the offer needing to respond if they wish to accept, or perhaps decline, the offer. The conditional commitment stems from the person who makes the offer being obligated to fulfil the promised action if the offer is accepted.

In Figure 3.24, the highlighted utterance from Kim to Kevin acts as a request for Kevin to respond with what he needs. It also acts as a conditional commitment for Kim to provide either a 2–3 page narrative or some existing slides, if Kevin responds. If Kevin fails to act to accept the offer or acts to decline the offer, then no obligation is placed on Kim. In this way, offers that are not accepted fail to satisfy our definition



From: Kimberly Watson <kimberly.watson@enron.com> **Sent:** Wed Jan 23 11:42:09 EST 2002
To: <a..howard@enron.com> ;
Cc:
Bcc:
Subject: RE: TransPecos and Sun Devil

```
Kevin,

For Trans Pecos, we have a couple of slides with a fact sheet prepared that we
used in the presentation with Stan a couple of weeks ago.  Would you like
those or are you looking for a 2-3 page narrative?  I'd be happy to provide
either.   What is the time frame - how quick?

Kim.

 -----Original Message-----
From:    Howard, Kevin A.
Sent:    Wednesday, January 23, 2002 1:34 PM
To:      Watson, Kimberly; Gadd, Eric
Subject:         TransPecos and Sun Devil

Eric and Kim:
I was hoping to send some information to a potential equity investor
(ArcLight) regarding both the TransPecos and Sun Devil transactions.  I have a
very brief presentation or detailed model - nothing in between.  I'm looking
for commercial overview, strategic rationale and summary economics.  Do you
have anything like this already prepared? If not, is there anyone who could
quickly pull this together?
Kevin
```

Figure 3.24: An email containing an offer.

of commitments, as outlined in section 3.1.3. Instead, at least initially, an offer acts more like a request, placing a conditional obligation on Kevin to respond if he wishes to accept or reject Kim's offer. Kim's original offer act then becomes a commitment if, and only if, Kevin responds and accepts her offer.

As the analysis of data from our annotation tasks highlights, the complexity within offers frequently leads to ambiguity about whether a particular offer conveys a request, a commitment or both. Deciding whether an utterance is phatic or actionable further complicates this decision. This complexity and ambiguity is widely acknowledged when we look into the speech act literature.

Offers are traditionally considered to have commissive force—e.g., (Austin, 1962; Searle, 1976; Leech, 1983; Bilbow, 2002). In our terminology, this equates to offers acting as commitments. In an email context, this seems reasonable given that when a sender offers something to a recipient, they are indeed conditionally committing themselves to future action, if the recipient accepts their offer. The situation is, however, more complex than this. As we have outlined above, a sender offering something also places some obligation on the recipient to respond to the offer, and in this way, offers can behave like requests.

While Austin and Searle considered offers to be commitments, subsequent studies have differed greatly over whether to categorise offers as commitments, requests—e.g., (Tsui, 1994)—or some hybrid category in between these—e.g., (Bach and Harnish, 1979; Hancher, 1979). As we noted earlier, Hancher argues for the creation of a separate, hybrid category which he calls Commissive Directives, a form of cooperative illocutionary act. Hancher reasons that "to offer something to someone is both to try to direct that person's behaviour, and also to commit oneself to a corresponding course of behavior"(Hancher, 1979, p. 6). He argues that these qualities make offers a form of hybrid speech act that combines a request and a commitment, with neither act dominating. According to this view, offering is treated as both a commitment and a request, and requires both participants to act. This coupling of speaker and recipient obligations is what defines offers as a cooperative illocutionary act.

This cooperative aspect of offers, and of the other acts discussed in this section, is what leads to the separation of the creation and activation of the associated obligation. While an offer sets up a future obligation for action, no obligation to act is triggered until a recipient accepts the offer. In this way, the action is conditioned on the recipient's acceptance of the offer. Because offers are a form of cooperative speech act, both parties are required to act before the obligation takes effect. As email is an asynchronous communication medium, this leads to the observed temporal separation of the creation and activation utterances associated with the obligation.

We consider that an email user may legitimately want to follow up on offers that are made. We recognise that it would be useful to be able to explicitly distinguish offers separately from other commitments, but we leave this to future work. When we asked annotators in our *Sentence 2* annotation experiment to mark commitments as conditional or unconditional, a key part of the distinction between promises and offers, the level of inter-annotator agreement was substantially lowered from $\kappa = 0.74$ to $\kappa = 0.65$.[14] We did not draw this distinction in any of our other annotated corpora of messages and text spans.

Instead, our guidance for interpreting offers is as follows:

> **Offers are commitments. Offers can also function as requests, which should be marked based on contextual interpretation.**

3.3.3.2 Invitations

Invitations are similar to offers, conveying aspects of both a request and a commitment. Consider the example below, drawn from an email about an upcoming seminar:

(3.87) *Join us for what should be a fun evening!*

Unlike his classification of offers, Searle considered invitations to be directives (Searle, 1976). There is clearly a request-like effect when an email sender invites a recipient

[14] Further detail about these finer-grained annotation results is provided in Section 3.2 and Appendix B.

Message	Commitment	No Commitment	Unsure	% Commitment	% No Commitment
1	23	26	4	43.40%	49.06%
2	36	16	1	67.92%	30.19%
3	34	17	2	64.15%	32.08%

Table 3.4: Agreement in 'crowd-sourced' classification of three email messages containing different forms of meeting related invitations.

to do something, as they are indeed trying to direct the recipient's behaviour. As we saw for offers, however, there is also a commitment associated with the act of inviting. If after inviting a recipient to a meeting, our email sender then refused to let them into the meeting room, the recipient would have reasonable grounds to object. This demonstrates the obligation that is also conveyed by an invitation that commits the email sender to future action.

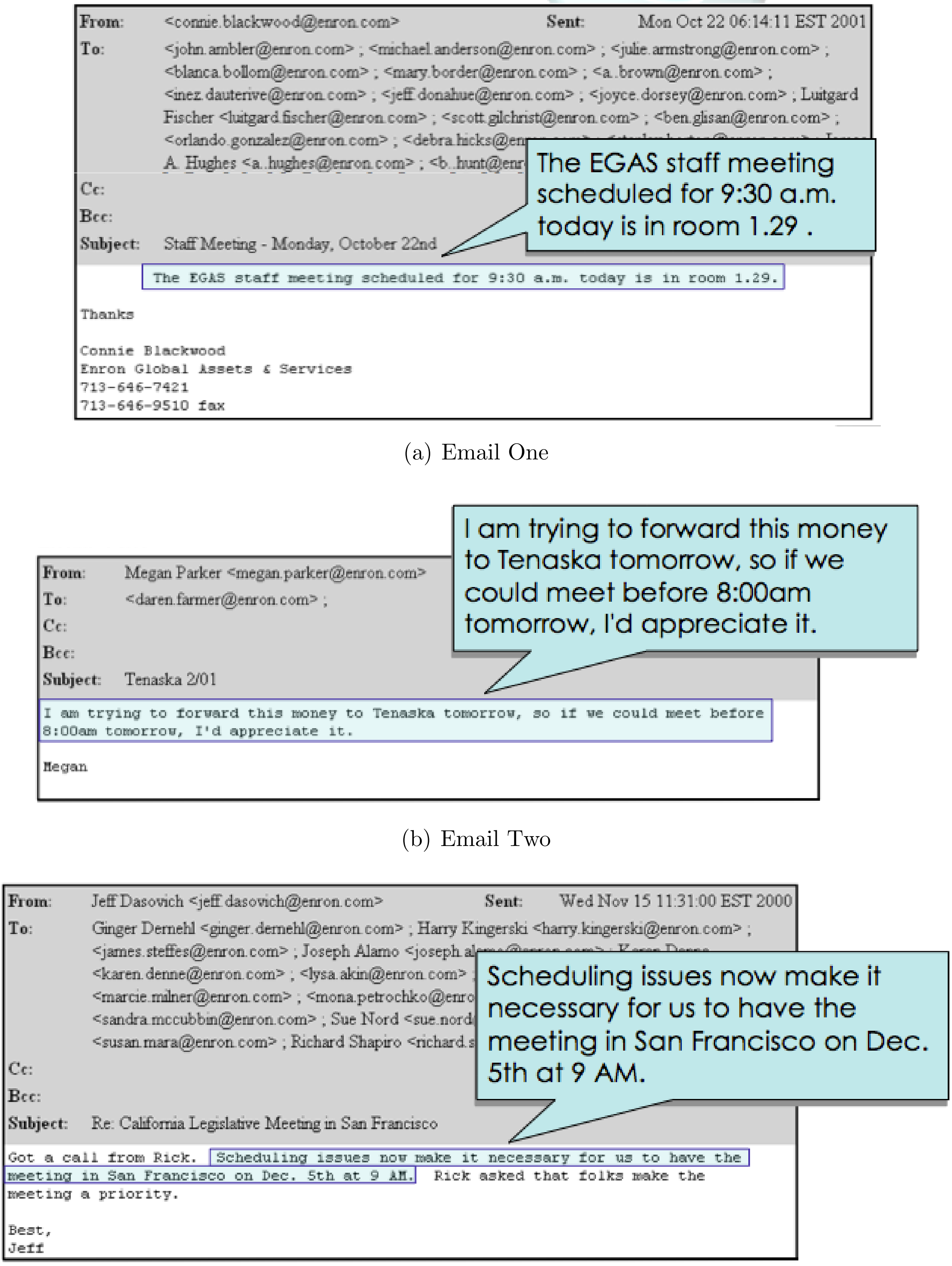
Given this, we would model the utterance shown in Example (3.87) as both a request and a commitment, as shown below:

(3.88) UR = ⟨Direct, *Join us for what should be a fun evening!*, R⟩
R = ⟨Action=Attend Seminar, Requestor=Michael Ellis,
Requestee=All Recipients⟩

(3.89) UC = ⟨Indirect, *Join us for what should be a fun evening!*, C⟩
C = ⟨Action=Allow Entry to Seminar, Committor=Michael
Ellis, Committee=Michael Ellis⟩

To further explore how people interpret invitations, we surveyed more than 50 people for their interpretation of commitments in three messages about meetings, shown in Figure 3.25.[15] The email in Figure 3.25(a) shows a clarification email about a meeting that appears to reinforce a previous invitation or request to attend a meeting. Figure 3.25(b) shows a message requesting or inviting the recipient to a meeting, and the third shows a reschedule of an existing meeting.

Table 3.4 summarises the results of our survey. The messages in Figure 3.25(b) and Figure 3.25(c) were both similarly interpreted, with 64–67% of respondents identifying that a commitment was conveyed.[16] One respondent commented that *"It depends if I was expected to be in the meeting"*, which again highlights that additional contextual understanding may be required to reliably interpret the invitation. The email in Figure 3.25(a) proved the most controversial, with our respondents roughly evenly split over whether the email created any commitment for the sender. Our respondents seemed to naturally focus more on the request component than the commitment for



From: <connie.blackwood@enron.com> Sent: Mon Oct 22 06:14:11 EST 2001
To: <john.ambler@enron.com> ; <michael.anderson@enron.com> ; <julie.armstrong@enron.com> ; <blanca.bollom@enron.com> ; <mary.border@enron.com> ; <a..brown@enron.com> ; <inez.dauterive@enron.com> ; <jeff.donahue@enron.com> ; <joyce.dorsey@enron.com> ; Luitgard Fischer <luitgard.fischer@enron.com> ; <scott.gilchrist@enron.com> ; <ben.glisan@enron.com> ; <orlando.gonzalez@enron.com> ; <debra.hicks@en
A. Hughes <a..hughes@enron.com> ; <b..hunt@enr
Cc:
Bcc:
Subject: Staff Meeting - Monday, October 22nd

```
        The EGAS staff meeting scheduled for 9:30 a.m. today is in room 1.29.

Thanks

Connie Blackwood
Enron Global Assets & Services
713-646-7421
713-646-9510 fax
```

The EGAS staff meeting scheduled for 9:30 a.m. today is in room 1.29 .

(a) Email One

From: Megan Parker <megan.parker@enron.com>
To: <daren.farmer@enron.com> ;
Cc:
Bcc:
Subject: Tenaska 2/01

```
I am trying to forward this money to Tenaska tomorrow, so if we could meet before
8:00am tomorrow, I'd appreciate it.

Megan
```

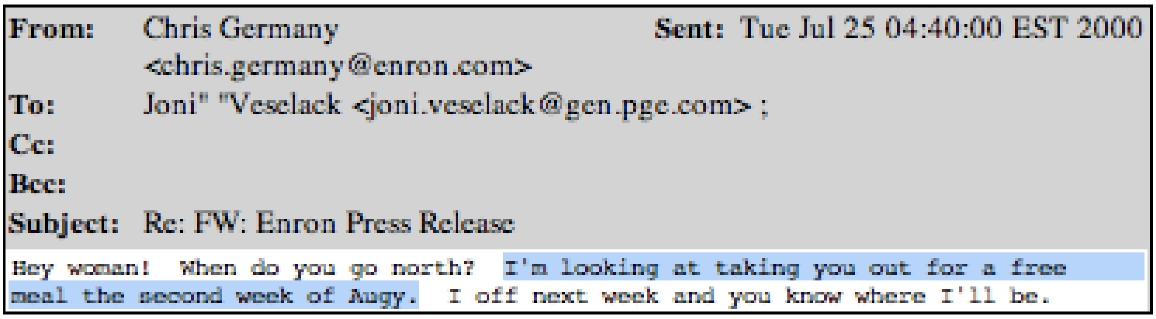
(b) Email Two

From: Jeff Dasovich <jeff.dasovich@enron.com> Sent: Wed Nov 15 11:31:00 EST 2000
To: Ginger Dernehl <ginger.dernehl@enron.com> ; Harry Kingerski <harry.kingerski@enron.com> ; <james.steffes@enron.com> ; Joseph Alamo <joseph.al
<karen.denne@enron.com> ; <lysa.akin@enron.com> ;
<marcie.milner@enron.com> ; <mona.petrochko@enro
<sandra.mccubbin@enron.com> ; Sue Nord <sue.nord
<susan.mara@enron.com> ; Richard Shapiro <richard.s
Cc:
Bcc:
Subject: Re: California Legislative Meeting in San Francisco

```
Got a call from Rick.  Scheduling issues now make it necessary for us to have the
meeting in San Francisco on Dec. 5th at 9 AM.  Rick asked that folks make the
meeting a priority.

Best,
Jeff
```

(c) Email Three

Figure 3.25: A collection of email messages that contain meeting-related commitments and requests.



From: Chris Germany <chris.germany@enron.com> **Sent:** Tue Jul 25 04:40:00 EST 2000
To: Joni" "Veselack <joni.veselack@gen.pge.com> ;
Cc:
Bcc:
Subject: Re: FW: Enron Press Release

```
Hey woman!  When do you go north?  I'm looking at taking you out for a free
meal the second week of Augy.  I off next week and you know where I'll be.
```

Figure 3.26: An invitation that functions as a commitment and, to a weaker extent, a request.

this message.

As should be clear by now, invitations can be complex and subjective to interpret. In some cases, including for the email shown in Figure 3.26, our human annotators marked invitations as both a request and a commitment. They also commented, however, that, because the highlighted utterance was preceded by an explicit request, the commissive force of the invitation was stronger than the request component. In our corpus data, very few invitations or offers were marked as a commitment without the request also being identified. One exception is the following example:

(3.90) *If you had only afternoon available, I can meet with you.*

The reverse, however, where only the request is marked, occurred frequently. This suggests that the directive force is frequently more dominant and relevant for an email recipient, which correlates with Searle's original classification of invitations as requests. It is also consistent with the responses from our survey for the messages in Figure 3.25.

A final remark about invitations is that responses to an invitation can also carry commissive force, and thus can themselves be classed as commitments. Take the concrete example of an invitation to a meeting, a very common form of invitation in workplace email. An acceptance to this invitation both obligates the person extending the invitation to allow the recipient to attend the meeting and obligates the responder to attend.

In summary, we provide the following guidance for interpreting invitations:

> **Invitations can function as both a request and a commitment; we leave it to contextual interpretation to decide which obligation acts are conveyed.**

[15] More detail about this survey is included in Section 3.3.9.

[16] There are, of course, also potential requests conveyed, but we asked respondents to focus on the commitments, if any, for these specific messages.

3.3.3.3 Threats

Threats share many properties with offers and invitations, including the separation of creation and activation of obligation. An example is shown below:

(3.91) *I'll fire you if you don't finish that work today.*

A threat is a form of conditional commitment, and, like an offer, tends to couple a request and a commitment together. The undesirability of the action from the point-of-view of the recipient is the key functional difference between an offer and a threat. This is, for example, the primary distinction between the threat in Example (3.92) from the offer in Example (3.93).

(3.92) *If you do not help me, I will fire you.*

(3.93) *If you need assistance, I'd be happy to help.*

This distinction can, however, be blurred when other intentions are delivered in the form of threats for purposes of humour, as in the following utterance:

(3.94) *If you keep working so hard, I'll be forced to give you a promotion.*

In our analysis of messages from the Enron corpus, threats are most commonly used as instruments of humour or irony. Even where threats are earnestly conveyed, they do not usually represent commitments that a recipient would want to follow up on; it is difficult to imagine an email recipient voluntarily seeking the execution of a threat that has been made against them. Our guidance for interpreting threats is, therefore, simple:

> **Threats are not actionable commitments in workplace communication. They frequently function as phatic commitments.**

3.3.3.4 Other Obligation Acts

The same separation of creation and activation of obligation that we have seen with offers and invitations can also be observed for requests that do not have a duality of also being commitments. An example of an utterance that creates but does not activate a request obligation is:

(3.95) *When I send you my edits, please post them to the website.*

The utterance in Example (3.95) creates an obligation for the recipient to act, with the action conditioned on further input from the sender. Later, when the sender sends through a second message with the required information and an utterance such as *Here are my edits*, the obligation is then activated for the recipient. As for the commitment example, the utterance *Here are my edits* does not create the obligation for the recipient, rather it triggers or activates the obligation inherent in the initial conditional request. On this basis, we consider the initial utterance that creates the obligation to be a request, but not the activating utterance.

Process instructions are another class of request that can regularly exhibit such separation of creation and activation of obligation. We discuss process instruction requests further in Section 3.3.5.

3.3.4 Transitive and Direct Obligation Acts

Obligation acts, particularly in a written medium such as email, frequently occur as both DIRECT ACTS—authored and sent by the sender of the message in which they appear—and TRANSITIVE ACTS—authored or uttered by a third-party. One commonly occurring type of transitive acts in email are REPORTED SPEECH ACTS, which include FORWARDED ACTS, speech acts that occur within forwarded content in an email message, as a specific sub-type.

Reported requests and commitment can be particularly sensitive to the context for their interpretation. Some reported requests, for example, clearly function as requests, as in the utterance in Example (3.96).

(3.96) *Paul asked if you could put together a summary of your accomplishments in an email.*

Other reported requests, however, do not impose an obligation on the recipient, such as the utterance shown in Example (3.97).

(3.97) *Sorry for the delay; Paul requested your prize to be sent out late December.*

Reported speech can similarly include commitments, as in Example (3.98).

(3.98) *Paul said that he would attend.*

In his influential work, Clark (1997) argues for a notion of layering in language use that is useful for analysing transitive speech acts in email.

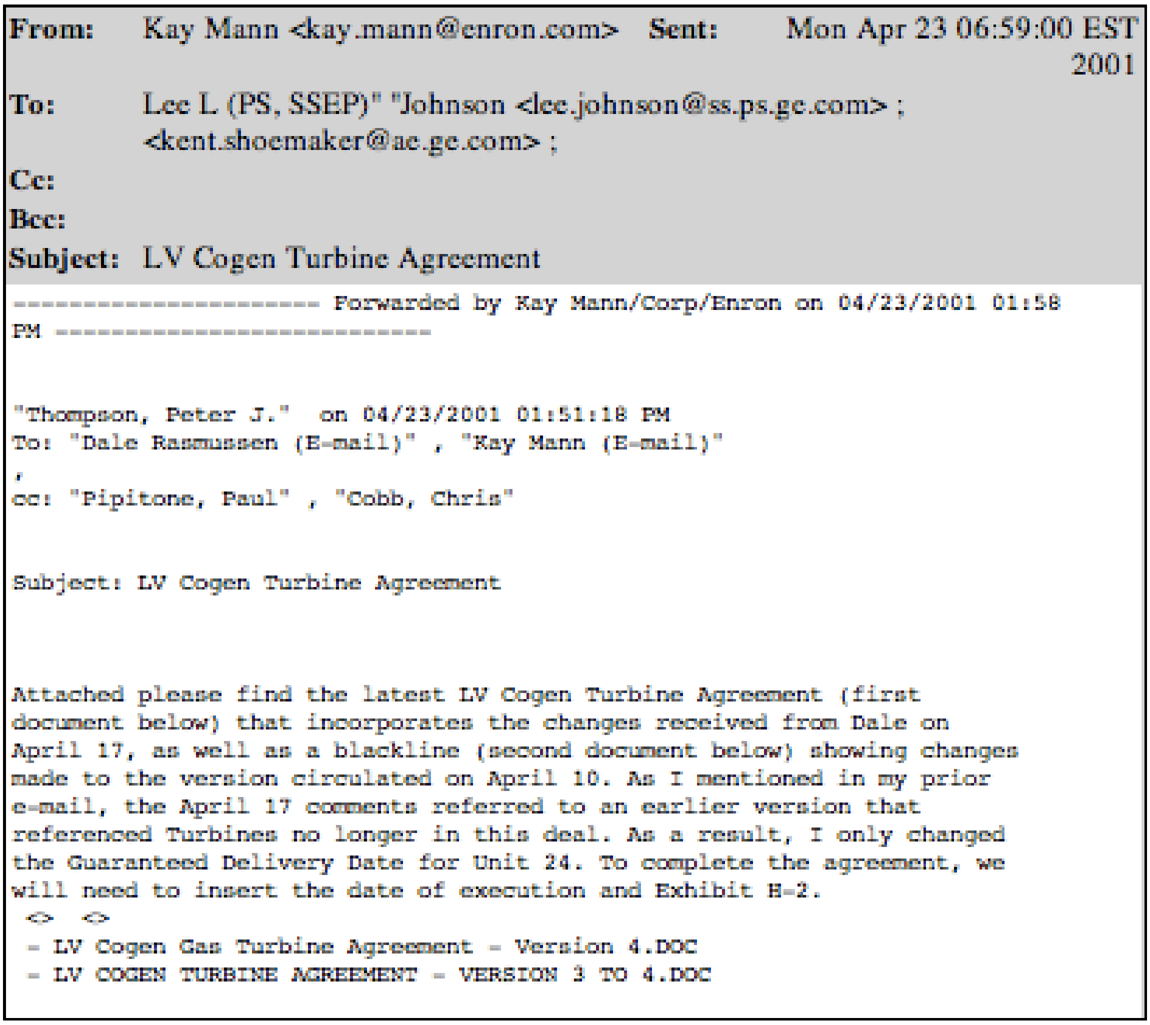
3.3.4.1 Layers in Language

People who take on one of the roles we have identified within email conversations in Section 3.1.4 are either participants or observers to a particular email communication. There may, however, be additional agents who contribute to the conversation content at different places and times, perhaps without having knowledge or visibility of the actual email communication. In our context of workplace email, an example of such agents include people whose speech acts are included within an email conversation as reported speech. An example would be the inclusion of an utterance like:

(3.99) *Robert said that he would attend.*

In this case, through his reported speech act, Robert is a participant in the conversation without his knowledge.

Clark's notion of layers in language use is much broader than our focus here (Clark, 1997), but his model is useful for our analysis of transitive speech acts in email. Each of Clark's layers is specified by its domain, notably who and what the participants are in that domain. Layers are "like theater stages built one on top of another" (Clark, 1997, p. 16). Clark posits that LAYER 1 represents the actual world, and is the primary layer of any conversation. Although not directly addressed by Clark, such a characterisation



From: Kay Mann <kay.mann@enron.com> **Sent:** Mon Apr 23 06:59:00 EST 2001

To: Lee L (PS, SSEP)" "Johnson <lee.johnson@ss.ps.ge.com> ;
<kent.shoemaker@ae.ge.com> ;

Cc:

Bcc:

Subject: LV Cogen Turbine Agreement

```
---------------------- Forwarded by Kay Mann/Corp/Enron on 04/23/2001 01:58
PM ---------------------------


"Thompson, Peter J."  on 04/23/2001 01:51:18 PM
To: "Dale Rasmussen (E-mail)" , "Kay Mann (E-mail)"
,
cc: "Pipitone, Paul" , "Cobb, Chris"


Subject: LV Cogen Turbine Agreement



Attached please find the latest LV Cogen Turbine Agreement (first
document below) that incorporates the changes received from Dale on
April 17, as well as a blackline (second document below) showing changes
made to the version circulated on April 10. As I mentioned in my prior
e-mail, the April 17 comments referred to an earlier version that
referenced Turbines no longer in this deal. As a result, I only changed
the Guaranteed Delivery Date for Unit 24. To complete the agreement, we
will need to insert the date of execution and Exhibit H-2.
 <>  <>
 - LV Cogen Gas Turbine Agreement - Version 4.DOC
 - LV COGEN TURBINE AGREEMENT - VERSION 3 TO 4.DOC
```

Figure 3.27: A forwarded message that contains an attachment request that may or may not place obligation for action on the current recipient, Lee Johnson.

equally applies to email conversations. Layer 1 is where the participants write as and are addressed as themselves, and is where actual people do actual things.

In the case of reported speech, we require the introduction of at least one more layer: the existing layer for the email sender and recipients to communicate, and another for the world in which the person whose speech act is being reported actually produced their utterance (outside of the email). Requiring the existence of this layer 2 is therefore an indicator of reported speech acts.

3.3.4.2 Reported Requests and Commitments

Reported speech acts, such as the two shown in Examples (3.96) and (3.97), are uttered by a third party but reported by the sender, in either the original utterer's or the sender's own paraphrasing words.

Email messages are frequently forwarded to other recipients in the workplace. In the Enron email collection, surface-level cues suggest that approximately 11% of emails are forwarded messages. The forwarded acts that occur within forwarded content are usually written by someone other than the current email sender, and are included in the current message through the forwarding function of the sender's email client. Forwarded acts can be thus be thought of as direct quotations from the original author. In this way, forwarded utterances act as a special case of reported speech within the context of email communication.

Requests and commitments that occur in forwarded content are a source of frequent disagreement among our human annotators. Consider the email in Figure 3.27. Is Lee Johnson, the recipient of this forwarded message, obligated to do anything with the included content and attached documents? As our guidance makes clear, we leave these judgements to the in-context interpretation of our annotators, rather than trying to provide overly prescriptive instructions. Importantly, however, transitive acts such as reported speech are clearly another source of complexity in how requests and commitments function in email.

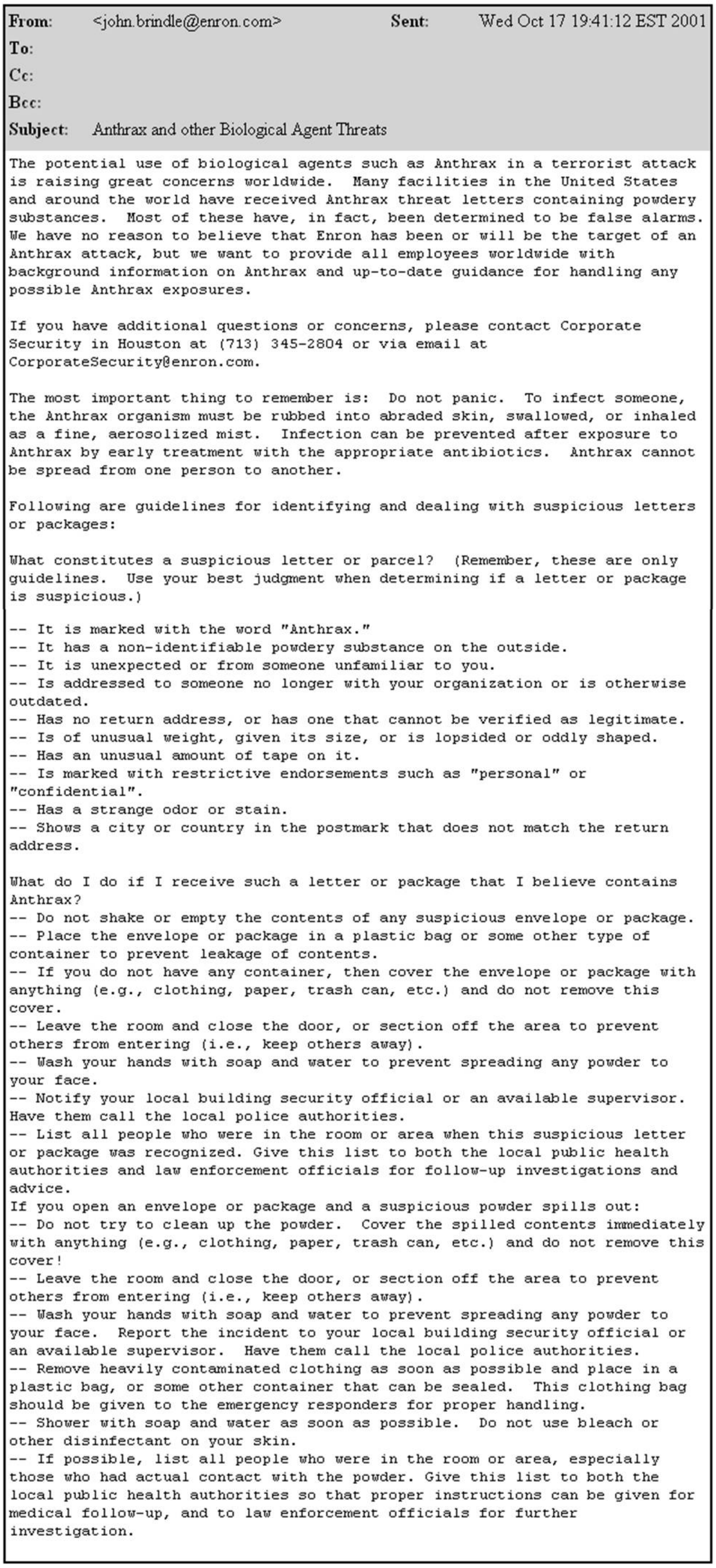
Our guidance for interpreting reported requests and commitments and related transitive acts is:

> **Some reported requests and commitments represent obligation acts; use the context to decide whether a reported act is actionable. Requests and commitments as forwarded acts are usually not actionable.**

3.3.5 Processes and Instructions

Email messages are frequently used to communicate instructions or descriptions of processes. Sometimes these instructions are of the kind that one might 'file for later use' and may never actually execute, such as in the email shown in Figure 3.28; at other times, instructions are intended to be executed more promptly, as in the email shown in Figure 3.29.

Our analysis of annotator disagreements about the status of instructions as requests suggests that several attributes influence whether an instruction should be marked as a request. One important factor is how soon after the message is received that action is expected. An instruction that should be followed in the near future would be more likely to be marked as a request. Another factor is the likelihood of the situation arising that would lead to the execution of the described instructions. The instructions at the bottom of the email in Figure 3.28 include:



From: <john.brindle@enron.com> **Sent:** Wed Oct 17 19:41:12 EST 2001

To:

Cc:

Bcc:

Subject: Anthrax and other Biological Agent Threats

The potential use of biological agents such as Anthrax in a terrorist attack is raising great concerns worldwide. Many facilities in the United States and around the world have received Anthrax threat letters containing powdery substances. Most of these have, in fact, been determined to be false alarms. We have no reason to believe that Enron has been or will be the target of an Anthrax attack, but we want to provide all employees worldwide with background information on Anthrax and up-to-date guidance for handling any possible Anthrax exposures.

If you have additional questions or concerns, please contact Corporate Security in Houston at (713) 345-2804 or via email at CorporateSecurity@enron.com.

The most important thing to remember is: Do not panic. To infect someone, the Anthrax organism must be rubbed into abraded skin, swallowed, or inhaled as a fine, aerosolized mist. Infection can be prevented after exposure to Anthrax by early treatment with the appropriate antibiotics. Anthrax cannot be spread from one person to another.

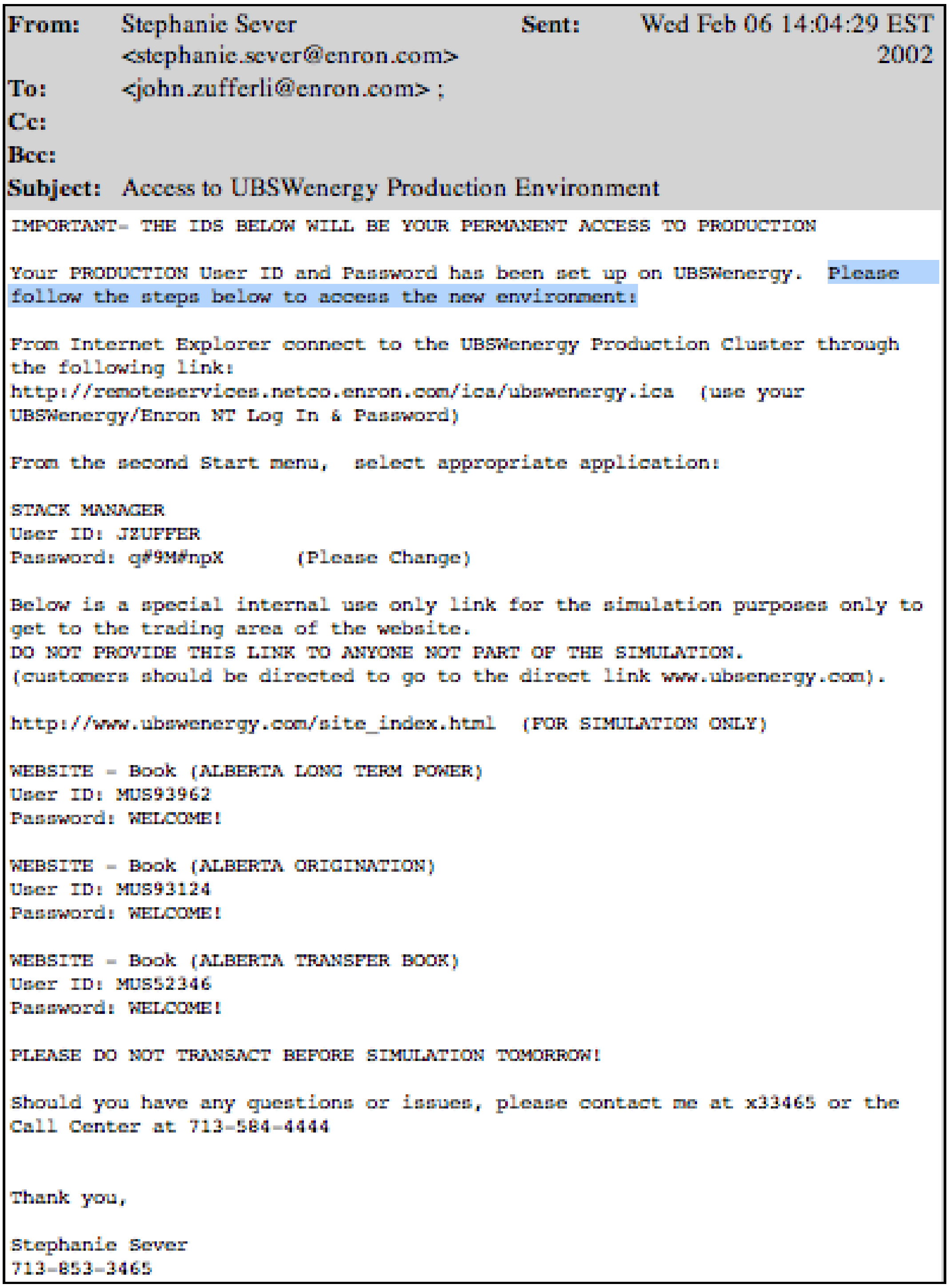
Following are guidelines for identifying and dealing with suspicious letters or packages:

What constitutes a suspicious letter or parcel? (Remember, these are only guidelines. Use your best judgment when determining if a letter or package is suspicious.)

-- It is marked with the word "Anthrax."
-- It has a non-identifiable powdery substance on the outside.
-- It is unexpected or from someone unfamiliar to you.
-- Is addressed to someone no longer with your organization or is otherwise outdated.
-- Has no return address, or has one that cannot be verified as legitimate.
-- Is of unusual weight, given its size, or is lopsided or oddly shaped.
-- Has an unusual amount of tape on it.
-- Is marked with restrictive endorsements such as "personal" or "confidential".
-- Has a strange odor or stain.
-- Shows a city or country in the postmark that does not match the return address.

What do I do if I receive such a letter or package that I believe contains Anthrax?
-- Do not shake or empty the contents of any suspicious envelope or package.
-- Place the envelope or package in a plastic bag or some other type of container to prevent leakage of contents.
-- If you do not have any container, then cover the envelope or package with anything (e.g., clothing, paper, trash can, etc.) and do not remove this cover.
-- Leave the room and close the door, or section off the area to prevent others from entering (i.e., keep others away).
-- Wash your hands with soap and water to prevent spreading any powder to your face.
-- Notify your local building security official or an available supervisor. Have them call the local police authorities.
-- List all people who were in the room or area when this suspicious letter or package was recognized. Give this list to both the local public health authorities and law enforcement officials for follow-up investigations and advice.
If you open an envelope or package and a suspicious powder spills out:
-- Do not try to clean up the powder. Cover the spilled contents immediately with anything (e.g., clothing, paper, trash can, etc.) and do not remove this cover!
-- Leave the room and close the door, or section off the area to prevent others from entering (i.e., keep others away).
-- Wash your hands with soap and water to prevent spreading any powder to your face. Report the incident to your local building security official or an available supervisor. Have them call the local police authorities.
-- Remove heavily contaminated clothing as soon as possible and place in a plastic bag, or some other container that can be sealed. This clothing bag should be given to the emergency responders for proper handling.
-- Shower with soap and water as soon as possible. Do not use bleach or other disinfectant on your skin.
-- If possible, list all people who were in the room or area, especially those who had actual contact with the powder. Give this list to both the local public health authorities so that proper instructions can be given for medical follow-up, and to law enforcement officials for further investigation.

Figure 3.28: An email containing process instructions to be filed for possible later use.



From: Stephanie Sever <stephanie.sever@enron.com>
Sent: Wed Feb 06 14:04:29 EST 2002
To: <john.zufferli@enron.com> ;
Cc:
Bcc:
Subject: Access to UBSWenergy Production Environment

```
IMPORTANT- THE IDS BELOW WILL BE YOUR PERMANENT ACCESS TO PRODUCTION

Your PRODUCTION User ID and Password has been set up on UBSWenergy.  Please
follow the steps below to access the new environment:

From Internet Explorer connect to the UBSWenergy Production Cluster through
the following link:
http://remoteservices.netco.enron.com/ica/ubswenergy.ica  (use your
UBSWenergy/Enron NT Log In & Password)

From the second Start menu,  select appropriate application:

STACK MANAGER
User ID: JZUFFER
Password: q#9M#npX      (Please Change)

Below is a special internal use only link for the simulation purposes only to
get to the trading area of the website.
DO NOT PROVIDE THIS LINK TO ANYONE NOT PART OF THE SIMULATION.
(customers should be directed to go to the direct link www.ubsenergy.com).

http://www.ubswenergy.com/site_index.html  (FOR SIMULATION ONLY)

WEBSITE - Book (ALBERTA LONG TERM POWER)
User ID: MUS93962
Password: WELCOME!

WEBSITE - Book (ALBERTA ORIGINATION)
User ID: MUS93124
Password: WELCOME!

WEBSITE - Book (ALBERTA TRANSFER BOOK)
User ID: MUS52346
Password: WELCOME!

PLEASE DO NOT TRANSACT BEFORE SIMULATION TOMORROW!

Should you have any questions or issues, please contact me at x33465 or the
Call Center at 713-584-4444


Thank you,

Stephanie Sever
713-853-3465
```

Figure 3.29: An email containing process instructions to be executed upon reading.

(3.100) *Wash your hands with soap and water to prevent spreading any powder to your face.*

(3.101) *Do not try to clean up the powder.*

(3.102) *Remove heavily contaminated clothing as soon as possible and place in a plastic bag, or some other container that can be sealed.*

Utterance such as these have a low probability of being actioned. Our human annotators were unlikely to mark these as a requests. Instructions for such low-probability events function more as information to be stored and referenced if required. In contrast, the instructions in Figure 3.29 require action from the recipient; they are instructions to be followed unconditionally, rather than being instructions for a hypothetical and unlikely situation.

Clark's concept of layering in language use, that we discussed in Section 3.3.4.1, neatly allows us to account for complex examples of instructions in email text. Instructions or descriptions of processes can differ in terms of which layer they occur in, which affects their interpretation and intention. Instructions that are addressed to the current recipient and are intended to be actioned "here and now" would be considered as part of Layer 1. More hypothetical instructions, or those filed for future use, could be considered to occur in a separate layer (e.g., a Layer 2). Within this analysis, it is only instructions or processes that occur within Layer 1 that would be marked as requests.

After careful consideration, we consider that:

> **Specific instances of process instructions are considered requests if they are intended to be actioned now and the local context supports this interpretation.**

3.3.6 Obligations for Inaction

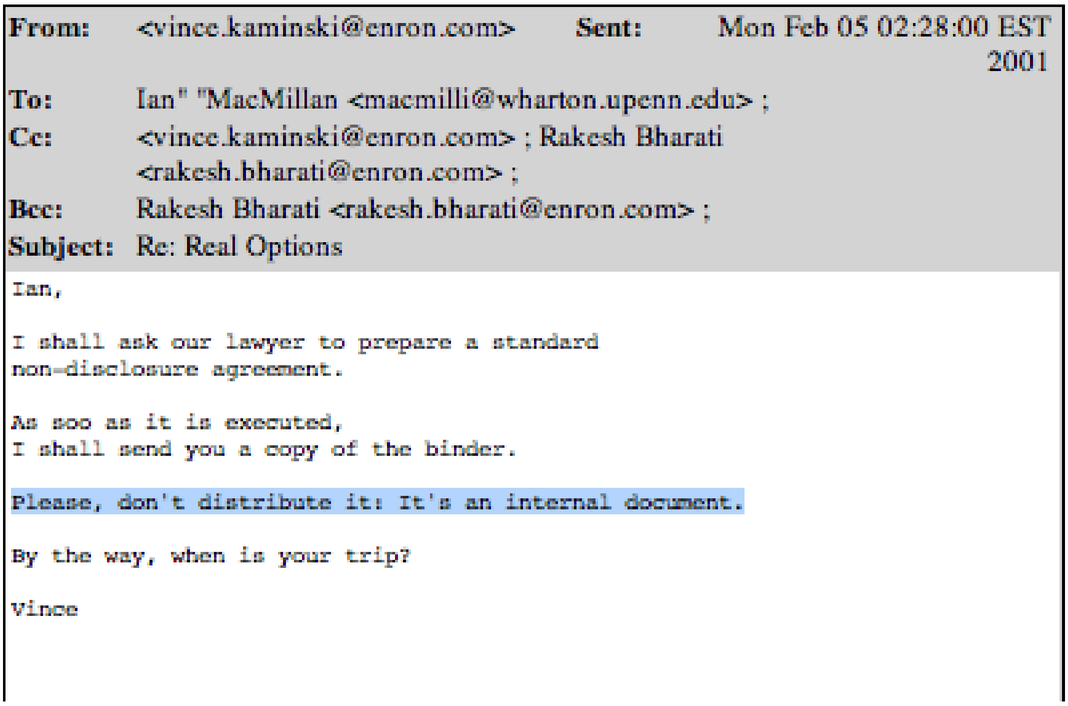
Obligations for inaction involve requests or commitments that ask or promise *not* to take some action. Requests for inaction are sometimes called PROHIBITIVES (Sadock and Zwicky, 1985), and frequently occur as utterances that either prohibit action, as in Example (3.103), or request negated action, as in Example (3.104), which is extracted from the email in Figure 3.30.

(3.103) *You may not charge this to your credit card.*

(3.104) *Please, don't distribute it*

Commitments to inaction are similar: they place obligation on the committee to *not* perform some future action. An example commitment to inaction is shown in Example (3.105):

(3.105) *I won't send you any further updates.*



From: <vince.kaminski@enron.com> **Sent:** Mon Feb 05 02:28:00 EST 2001
To: Ian" "MacMillan <macmilli@wharton.upenn.edu> ;
Cc: <vince.kaminski@enron.com> ; Rakesh Bharati <rakesh.bharati@enron.com> ;
Bcc: Rakesh Bharati <rakesh.bharati@enron.com> ;
Subject: Re: Real Options

```
Ian,

I shall ask our lawyer to prepare a standard
non-disclosure agreement.

As soo as it is executed,
I shall send you a copy of the binder.

Please, don't distribute it: It's an internal document.

By the way, when is your trip?

Vince
```

Figure 3.30: An email containing a request for inaction: *Please, don't distribute it.*

By definition, requests for and commitments to inaction do not require action, so one would not expect such utterances to result in new tasks being created in either the recipient's or the sender's task list. As a result, definitions based on the suitability of an action for inclusion in a recipient's task list fail to account for requests and commitments for inaction. Clearly, however, such requests and commitments do place an obligation on the recipient or the sender.[17]

It is worth noting, however, that utterances that merely request or commit to action using a negated surface form, such as the utterances in Example (3.106) and Example (3.107), are requests for action rather than requests for inaction.

(3.106) *Don't forget to send me your comments*

(3.107) *I won't forget to send you my comments*

The utterance in Example (3.106) is simply an alternate realisation of the utterance *Please send me your comments*. Similarly, the utterance in Example (3.107) is an alternate realisation of *I will send you my comments*. The contrasting use of negation

[17] Our inclusion of requests for inaction as requests stems from consideration of the application end goals where requests for inaction should be identified. Note that this differs from our original stance taken in (Lampert, Dale, and Paris, 2008b), which was to not interpret requests for inaction as obligation acts. Our change in stance is due to broader thinking about the requirements for different end applications, as discussed in Section 3.1.8.

illustrates further complexity in the relationship between the surface text of an uttered request or commitment and the underlying speech act.

Indeed, requests for inaction need *not* be expressed in a negated form, as we see in the following examples that each contain a positively phrased request for inaction:

(3.108) *Please keep to yourself at this point.*

(3.109) *I have sent a copy to Tom Zenner, but please keep this very confidential.*

Our guidance for interpreting obligations for inaction is:

> **Requests for inaction and commitments to inaction are requests and commitments respectively.**

3.3.7 Justification, Elaboration and Other Supporting Text

As Sandra Thompson and Bill Mann noted in their work on Rhetorical Structure Theory, "for a written directive to succeed and convince us to comply with a request or accept an offer, there may be portions of text devoted to motivating us to comply and to letting us know how to comply" (Thompson and Mann, 1987, p85).

We commonly see such supporting information accompanying requests and commitments in email, perhaps being especially important given the lack of paralinguistic cues such as gesture and intonation in email communication. An example of a request being followed by a clause providing support in the form of justification is shown in Figure 3.31. In this case, Vince includes the phrase shown in Example (3.110) as a parenthetical justification for his request to Andreas to fix any mistakes in the material he has provided.

(3.110) *... (no spell checker on my laptop for AOL).*

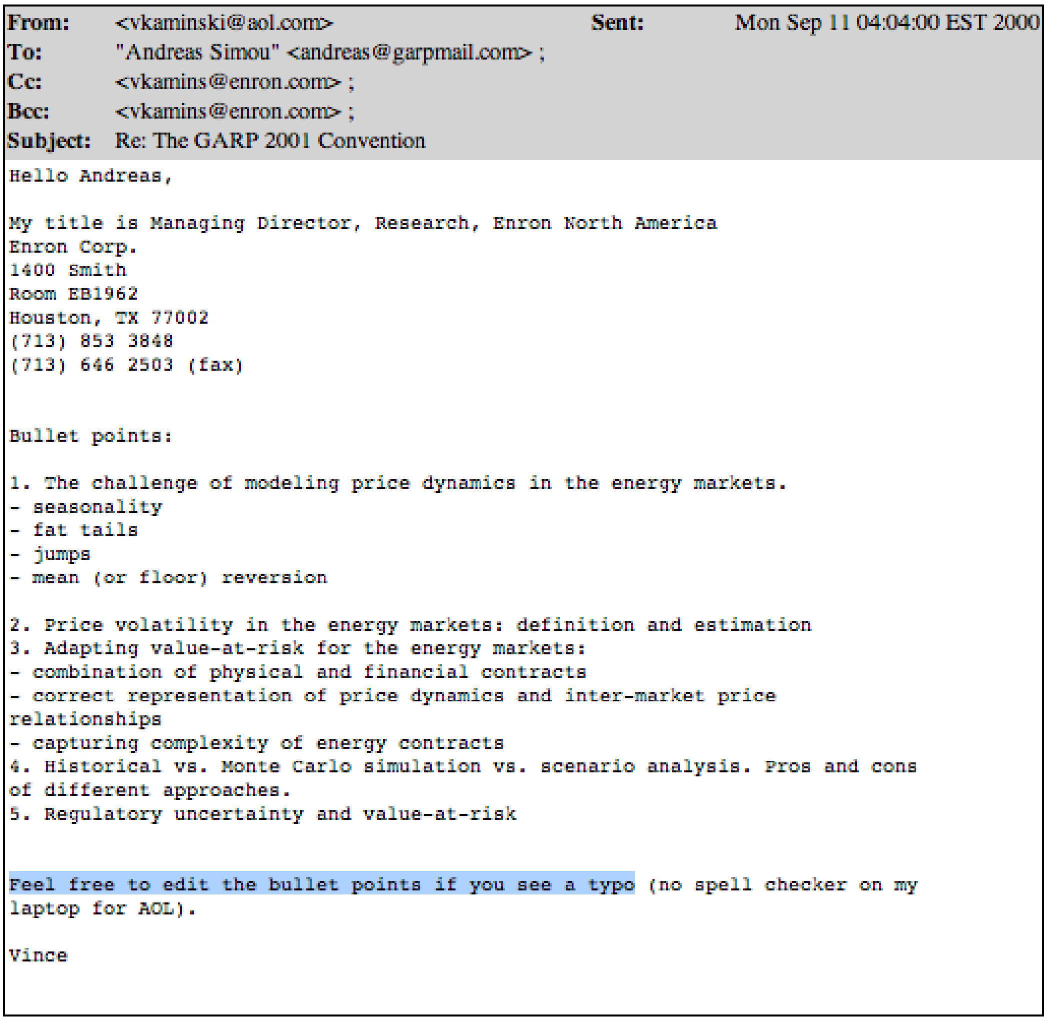
When identifying obligation acts, a common disagreement that arises is whether information that justifies, elaborates or otherwise supports the obligation act, such as that shown in Example (3.110) should be included.

The same style of supporting justification can also be observed with commitments, as shown in Figure 3.32. In this example, the negative commitment made by the sender, Carolyn, and shown in Example (3.111), is followed by a justification of why she cannot participate, as shown in Example (3.112).

(3.111) *I'm sorry I will be unable to participate*

(3.112) *this meeting conflicts with an IEP Board meeting*

As for the example from Figure 3.31, the justification occurs within the same sentence, and our annotators commented on ambiguity around whether to include the justification as part of the commitment.



From: <vkaminski@aol.com> Sent: Mon Sep 11 04:04:00 EST 2000
To: "Andreas Simou" <andreas@garpmail.com> ;
Cc: <vkamins@enron.com> ;
Bcc: <vkamins@enron.com> ;
Subject: Re: The GARP 2001 Convention

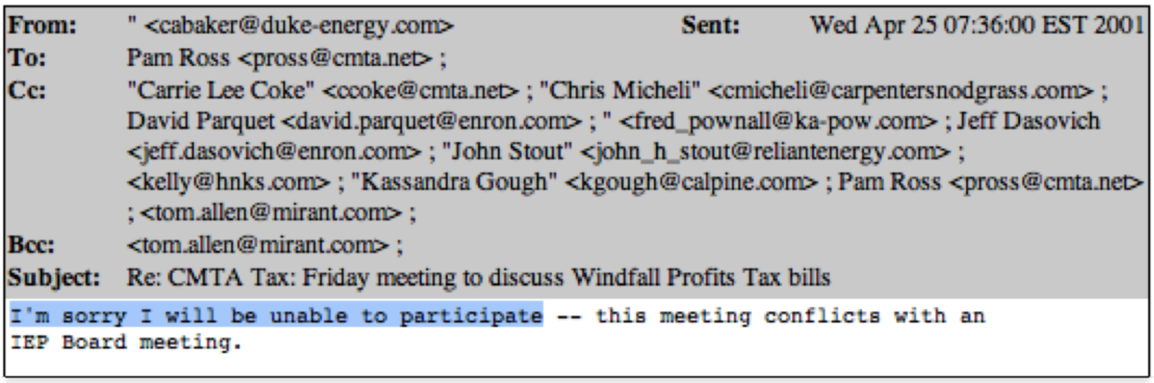
```
Hello Andreas,

My title is Managing Director, Research, Enron North America
Enron Corp.
1400 Smith
Room EB1962
Houston, TX 77002
(713) 853 3848
(713) 646 2503 (fax)


Bullet points:

1. The challenge of modeling price dynamics in the energy markets.
- seasonality
- fat tails
- jumps
- mean (or floor) reversion

2. Price volatility in the energy markets: definition and estimation
3. Adapting value-at-risk for the energy markets:
- combination of physical and financial contracts
- correct representation of price dynamics and inter-market price
relationships
- capturing complexity of energy contracts
4. Historical vs. Monte Carlo simulation vs. scenario analysis. Pros and cons
of different approaches.
5. Regulatory uncertainty and value-at-risk


Feel free to edit the bullet points if you see a typo (no spell checker on my
laptop for AOL).

Vince
```

Figure 3.31: An email message with a highlighted request, followed by a justification clause.

This complexity is not limited to justification phrases; utterances offering many forms of rhetorical support for a request or commitment may be found in our email datasets. Elaboration is another common form of supporting material that can accompany requests and commitments, as illustrated in Figure 3.33. In this case, the core request is:

(3.113) *I'm writing to urge you to donate the millions of dollars you made from selling Enron stock before the company declared bankruptcy*

This request is followed by several clauses of elaboration and justification of exactly what the request is:



From:	" <cabaker@duke-energy.com>	**Sent:**	Wed Apr 25 07:36:00 EST 2001
To:	Pam Ross <pross@cmta.net> ;		
Cc:	"Carrie Lee Coke" <ccoke@cmta.net> ; "Chris Micheli" <cmicheli@carpentersnodgrass.com> ; David Parquet <david.parquet@enron.com> ; " <fred_pownall@ka-pow.com> ; Jeff Dasovich <jeff.dasovich@enron.com> ; "John Stout" <john_h_stout@reliantenergy.com> ; <kelly@hnks.com> ; "Kassandra Gough" <kgough@calpine.com> ; Pam Ross <pross@cmta.net> ; <tom.allen@mirant.com> ;		
Bcc:	<tom.allen@mirant.com> ;		
Subject:	Re: CMTA Tax: Friday meeting to discuss Windfall Profits Tax bills		

```
I'm sorry I will be unable to participate -- this meeting conflicts with an
IEP Board meeting.
```

Figure 3.32: An email message with a highlighted negative commitment, followed by a justification clause.

(3.114) *... to funds, such as Enron Employee Transition Fund and REACH, that benefit the company's employees, who lost their retirement savings, and provide relief to low-income consumers in California, who can't afford to pay their energy bills*

This supporting material in Example (3.114) is not a request, and should not be marked as part of the request in Example (3.113).

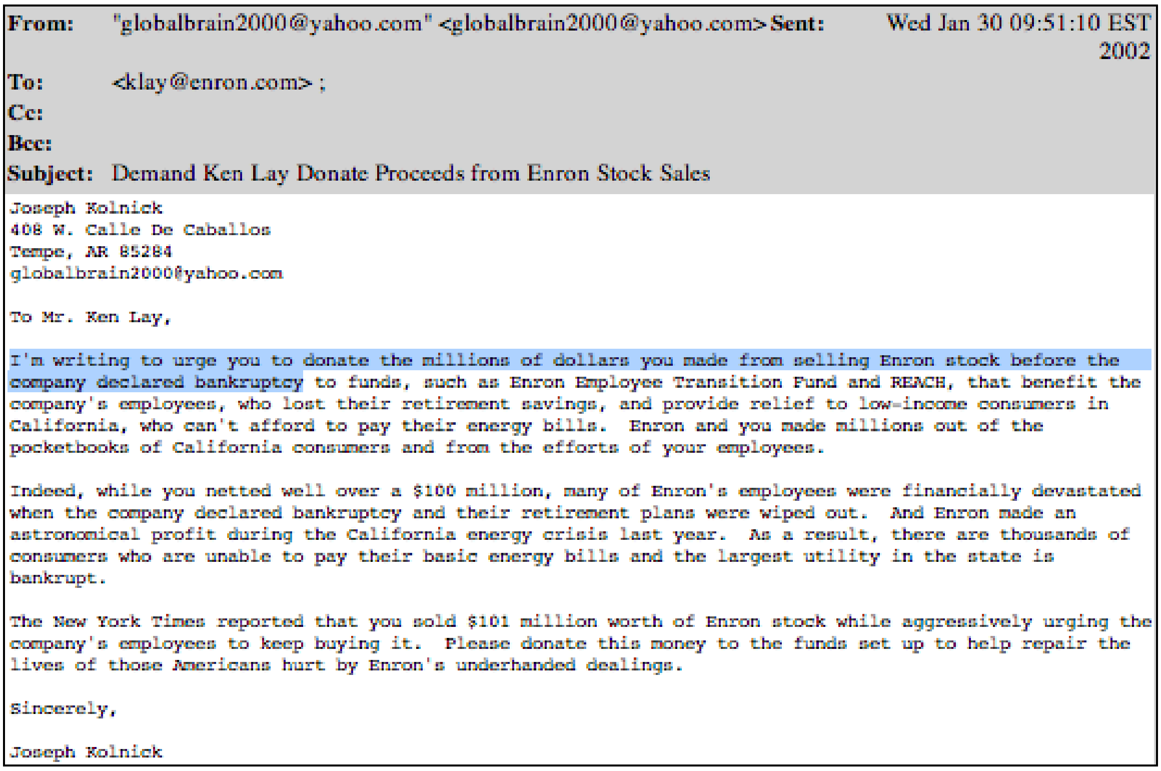
Remembering that our goal is to identify each and every realisation of a request or commitment, it can be particularly ambiguous whether a supporting utterance is functioning as an implicit request or commitment, or simply as supporting material for a more explicit realisation of the request or commitment elsewhere in the message. In Figure 3.32, for example, consider the utterance:

(3.115) *this meeting conflicts with an IEP Board meeting*

Had Carolyn Baker's message included only the justification utterance in Example (3.115), it seems likely that this would have been interpreted in isolation as an implicit negative commitment to the original request. As it stands in the actual context, however, the role of the utterance in Example (3.115) seems supporting, and thus it would not be included or identified as a commitment. The distinction between these supporting and implicit restatements of requests and commitments can, however, clearly be highly context sensitive.

Given the complexity, our guidance for interpreting supporting information is:

> **Supporting material does not represent part of an obligation act, where it exists only to elaborate, justify or otherwise support an uttered request or commitment.**



From: "globalbrain2000@yahoo.com" <globalbrain2000@yahoo.com> **Sent:** Wed Jan 30 09:51:10 EST 2002

To: <klay@enron.com> ;

Cc:

Bcc:

Subject: Demand Ken Lay Donate Proceeds from Enron Stock Sales

```
Joseph Kolnick
408 W. Calle De Caballos
Tempe, AR 85284
globalbrain2000@yahoo.com

To Mr. Ken Lay,

I'm writing to urge you to donate the millions of dollars you made from selling Enron stock before the
company declared bankruptcy to funds, such as Enron Employee Transition Fund and REACH, that benefit the
company's employees, who lost their retirement savings, and provide relief to low-income consumers in
California, who can't afford to pay their energy bills.  Enron and you made millions out of the
pocketbooks of California consumers and from the efforts of your employees.

Indeed, while you netted well over a $100 million, many of Enron's employees were financially devastated
when the company declared bankruptcy and their retirement plans were wiped out.  And Enron made an
astronomical profit during the California energy crisis last year.  As a result, there are thousands of
consumers who are unable to pay their basic energy bills and the largest utility in the state is
bankrupt.

The New York Times reported that you sold $101 million worth of Enron stock while aggressively urging the
company's employees to keep buying it.  Please donate this money to the funds set up to help repair the
lives of those Americans hurt by Enron's underhanded dealings.

Sincerely,

Joseph Kolnick
```

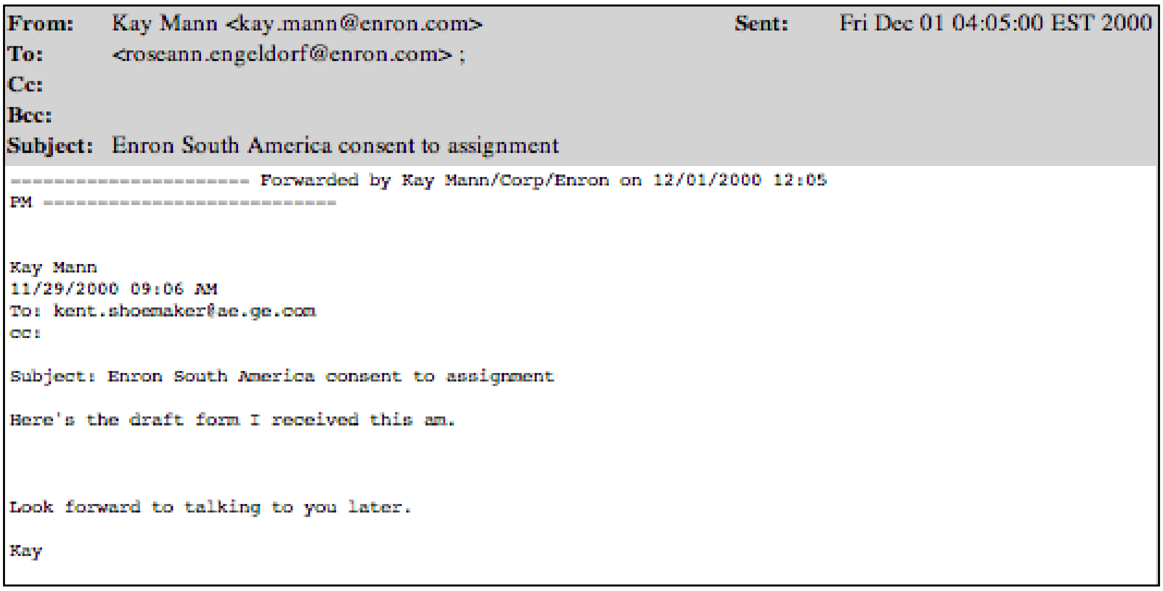
Figure 3.33: An email message with a highlighted request, followed by several clauses of elaboration.

3.3.8 The Effects of Message-Level Structure and Message Genre

The functional layout of content in an email message can have complex effects on the interpretation of requests and commitments. In this section, we briefly analyse how this influences the interpretation of requests and commitments. We return to this issue in more depth in Chapter 4.

Somewhat separate from the discourse structure of the message text, as we have touched on in our discussion of supporting content in Section 3.3.7, the body text of an email message also has a rich functional structure in terms of its layout and presentation. For example, the content of an email disclaimer is functionally different from the sender-authored content and from the quoted reply content automatically included from previous messages in the thread of conversation.

When identifying requests and commitments, only content from some of these message parts is relevant. If a request or commitment occurs in a replied-to message that is quoted within the current message, for example, then it is likely that this request or commitment was actionable for the *original* recipient, but not for a recipient of the current message. Similarly, we may wish not to identify requests in embedded advertising



```
From:    Kay Mann <kay.mann@enron.com>                Sent:    Fri Dec 01 04:05:00 EST 2000
To:      <roseann.engeldorf@enron.com> ;
Cc:
Bcc:
Subject: Enron South America consent to assignment
---------------------- Forwarded by Kay Mann/Corp/Enron on 12/01/2000 12:05
PM ---------------------------


Kay Mann
11/29/2000 09:06 AM
To: kent.shoemaker@ae.ge.com
cc:

Subject: Enron South America consent to assignment

Here's the draft form I received this am.



Look forward to talking to you later.

Kay
```

Figure 3.34: An email message with a forwarded attachment request that may or may not be actionable for Roseann Engeldorf, the current email recipient.

content given their likely low importance. The relevant characteristic of these requests and commitments that we wish to ignore lies in the specific part of the message in which they occur.

We refer to the different functional parts of a message as EMAIL ZONES, and the task of identifying the different functional parts as EMAIL ZONING. We discuss a categorisation of email zones and techniques for automatically classifying email zones in Section 4.4. For the purposes of our annotation tasks, and for our later automated work, we only consider requests and commitments in a subset of email zones.

Even with an agreed segmentation of an email message into zones and an agreed set of relevant zones whose content will be considered, however, there can still be ambiguity. The most frequently observed source of disagreement about zone relevance in our annotation experiments has stemmed from forwarded content, that is, content that is quoted from an earlier email message and forwarded to a different audience. In many cases, this content serves as context for the current sender's content; in some cases, however, forwards appear to be used as a form of delegation, as in Figure 3.34.

Our guidance on interpreting content from different functional parts of an email message is as follows:

> **Any requests or commitments in text that appears in the content of earlier messages should be ignored. Similarly, any requests or commitments in boilerplate email zones should be ignored.**[18]

[18] As we elaborate in Section 4.4, boilerplate zones include email signatures, legal disclaimers and

In a similar manner to the way we ignore content in certain email zones, we also consider that certain genres of messages should not be analysed for request and commitment content. The most obvious of these are spam messages. Other message genres that may be ignored include bulk marketing or promotional messages, automated notification emails and many forms of so-called 'organisational spam' or BACN messages, that do not generally contain actionable content. In general, however, we consider this problem to be best addressed through separate message classification that would exclude irrelevant message genres from obligation act processing.

3.3.9 Requests Related to Attached Documents

Interpreting utterances that make reference to reading, filing or otherwise acting on an attached document were a significant source of complexity that led to disagreement through some of our early annotation experiments. An example of such an utterance is *Please see attached*, as shown in the email message in Figure 3.35(a).

To help us understand how people interpret utterances that either implicitly or explicitly make reference to an attached document, we undertook a specific survey to gather a wide range of interpretations for a small number of email messages which contained a variety of forms of attachment-related requests. The setting for this survey was a 'Speed Paper' session in a conference setting.[19] The audience was polled for their interpretations of six email messages, presented on an overhead projector after being provided with some context and instructions for whose perspective to adopt, and the broader goals of our work. We have discussed three of these messages back in Section 3.3.3.2. The remaining three messages each contained an attached document and utterances that contained possible requests related to the attached document. These three messages are shown in Figure 3.35. The specific question posed for each of these three messages was *If you received this message, would you feel obligated to act, schedule future action or respond?*. More than fifty people submitted their responses in written form along with free-form comments about their interpretation.

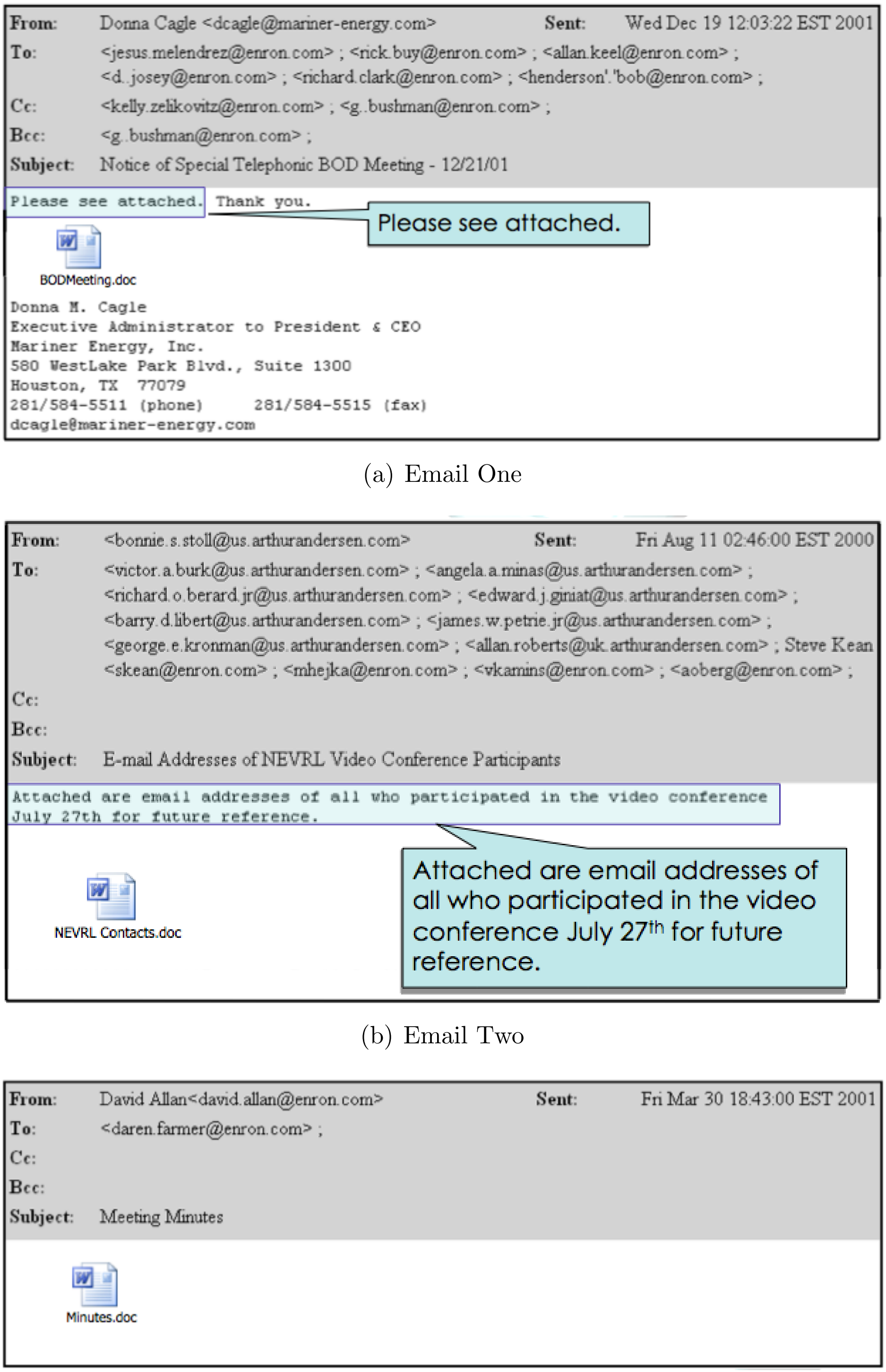
As illustrated, the email in Figure 3.35(a) contains an imperative utterance that references the attached document, the second email, Figure 3.35(b), contains a declarative utterance that references an attached document, and the third email, Figure 3.35(c), makes no explicit reference to the attached document, beyond the phrase *Meeting Minutes* in the subject text.

The results of this 'crowd-sourced' survey are shown in Table 3.5. There are several interesting observations that can be drawn from this activity. Firstly, it is clear that there is significant variation in interpretation across the three messages, with the best agreement (87%) being achieved for the message in Figure 3.35(b). Most respondents did not interpret the declarative utterance in this message as a request.[20] A few

automatically inserted advertising.

[19] The specific conference was the 2008 HCSNet (the Australian Research Council Network in Human Communication Science) Summerfest, held at the University of New South Wales.

[20] The number of responses for each message is not identical, as some participants failed to include



From: Donna Cagle <dcagle@mariner-energy.com> Sent: Wed Dec 19 12:03:22 EST 2001
To: <jesus.melendrez@enron.com> ; <rick.buy@enron.com> ; <allan.keel@enron.com> ;
<d..josey@enron.com> ; <richard.clark@enron.com> ; <henderson'.'bob@enron.com> ;
Cc: <kelly.zelikovitz@enron.com> ; <g..bushman@enron.com> ;
Bcc: <g..bushman@enron.com> ;
Subject: Notice of Special Telephonic BOD Meeting - 12/21/01

Please see attached. Thank you.

Please see attached.

BODMeeting.doc

Donna M. Cagle
Executive Administrator to President & CEO
Mariner Energy, Inc.
580 WestLake Park Blvd., Suite 1300
Houston, TX 77079
281/584-5511 (phone) 281/584-5515 (fax)
dcagle@mariner-energy.com

(a) Email One

From: <bonnie.s.stoll@us.arthurandersen.com> Sent: Fri Aug 11 02:46:00 EST 2000
To: <victor.a.burk@us.arthurandersen.com> ; <angela.a.minas@us.arthurandersen.com> ;
<richard.o.berard.jr@us.arthurandersen.com> ; <edward.j.giniat@us.arthurandersen.com> ;
<barry.d.libert@us.arthurandersen.com> ; <james.w.petrie.jr@us.arthurandersen.com> ;
<george.e.kronman@us.arthurandersen.com> ; <allan.roberts@uk.arthurandersen.com> ; Steve Kean
<skean@enron.com> ; <mhejka@enron.com> ; <vkamins@enron.com> ; <aoberg@enron.com> ;
Cc:
Bcc:
Subject: E-mail Addresses of NEVRL Video Conference Participants

Attached are email addresses of all who participated in the video conference July 27th for future reference.

Attached are email addresses of all who participated in the video conference July 27th for future reference.

NEVRL Contacts.doc

(b) Email Two

From: David Allan<david.allan@enron.com> Sent: Fri Mar 30 18:43:00 EST 2001
To: <daren.farmer@enron.com> ;
Cc:
Bcc:
Subject: Meeting Minutes

Minutes.doc

(c) Email Three

Figure 3.35: Three email messages with possible attachment-related requests, that were presented for interpretation in a conference setting. The form of the messages shown here is exactly the way these messages were presented for interpretation, one-at-a-time on an overhead projector.

a judgement for every message.

Message	Request	No Request	Unsure	% Request	% No Request
1	42	10	0	80.77%	19.23%
2	6	47	1	11.11%	87.04%
3	15	37	2	27.78%	68.52%

Table 3.5: A table showing agreement in our survey of interpretations for three email messages containing different forms of attachment requests.

respondents qualified their interpretation based on an assumed context of reading that made them *"too busy"* to interpret the utterance as a request, but noted that otherwise they might read it as a request. This reveals yet further complexity in the interpretation of requests: to some degree, the recipient may modify their interpretation based on their level of busyness.

A similar number of respondents (more than 80%) felt that the first message, with an imperative form utterance making reference to a different attached document, *did* represent a request. This was fairly universally interpreted as the request being a variation of *Please read the attachment*. This confirmed for us that, under the appropriate conditions, people do interpret explicit requests related to attachments as genuine requests.

Finally, opinions were divided on the third message without an explicit reference to the attachment. One person commented that they would mark it as a request *"if I'd been to the meeting"*, while another observed that it *"depends on events @ this meeting"*. Yet another respondent considered minutes to be a specific type of attachment that he would always interpret as a request, noting that *minutes need to be read and approved.* Clearly not all attached documents are equivalent in terms of how utterances that create obligations related to them will be interpreted.

In addition to the quantitative data and specific comments about some of the messages, there were some interesting general comments made about the task, including:

- *"Need contextual knowledge"*
- *"I found it difficult to make judgements. I think the on-going and historical relationship between communicators, and the importance of the topic would play major roles in these decisions."*
- *"Much harder to determine whether R or C than I thought it would be!"*
- *"I found it interesting in terms of the judgement required—i.e., to what degree or extent I would consider it a request/commitment."*
- *"Interesting, but I'd like to know the solution :-)"*

Collectively, these comments confirm that requests and commitments are more complex phenomena to identify than people intuitively expect. Partly this is because the interpretation of requests and commitments is deeply context-sensitive. This leads us to our discussion in Chapter 6 on in-context annotation, and the need for tools and processes that can capture more nuanced judgements about human language use: such concerns go beyond syntax and semantics to *pragmatic* aspects of language use, which explore the meaning of language in its real-world context. In Chapter 6, we explore the development of ecologically-valid tools that closely approximate the real-world environment in which the documents under examination are created and used.

Overall, it is clear that attachment-related requests are another complex phenomenon found in real-world email. Our observations and analysis from the additional survey data, coupled with observations from our annotation experiments, has helped us arrive at the following guidance for interpreting attachment-related requests:

> **Messages with attachments are requests to open, read, or act on the attached document(s) if, on the basis of the content of the email message, you believe you would feel obliged to open, read or otherwise act upon the attached file(s).**

3.4 Summary

As we noted at the beginning of this chapter, the task of identifying requests and commitments in email involves recognising and interpreting an email author's intentions. This makes our task an instance of the more general problem of recognising the speech acts behind an author's utterances. As with other intention recognition problems, our work has confirmed that the task of identifying obligation acts in email text can be highly context-sensitive.

Throughout this chapter, we have carefully defined what we consider to be (and not to be) requests and commitments in the context of workplace email. We have done this based on a theoretical linguistic understanding of how requests and commitments function coupled with explicit empirical evidence and observations largely derived from the series of manual annotation experiments that we described in Section 3.2. From the theoretical perspective, we have accounted for properties such as conditionality, and considered important factors such as the audience and addressivity within email text. The annotated data from our experiments supports this work, and reveals further complexity and ambiguity in the way people create and interpret requests and commitments. In Section 3.3, we identified and analysed the most common complex phenomena that we observed in obligation act usage. As part of the analysis of these phenomena, we have proposed principles of interpretation that define where the boundaries lie between obligation acts and other speech acts.

Specifically, we have provided guidance on how we interpret obligations in each of the following cases:

- phatic obligation acts;
- obligation acts that require ongoing action and one-time action;
- conversations where the creation and activation of the obligation is distributed across different messages;

- transitive obligation acts that are uttered by a third party and conveyed by an email sender, such as reported and forwarded obligation acts;
- processes and instructions;
- obligations for inaction;
- additional utterances that provide rhetorical support, such as justification or elaboration, for a request or a commitment;
- requests and commitments in different genres of messages and different functional parts of an email message, such as in forwarded, quoted reply or email signature content; and
- requests that make reference to reading, saving or otherwise acting on attached documents.

Together, these phenomena blur the boundaries between utterances that represent obligation acts and those that do not. Our observations and analysis reveal considerable complexity in the way that requests and commitments are exchanged and employed in real-world email. An important implication of this chapter is that the phenomena and features that we have identified should be explicitly considered before one attempts to either manually or automatically identify requests or commitments in email text. Our review of related literature suggests that this complexity has often been overlooked.

The theoretical and empirical observations and analysis in this chapter lay the foundation for the rest of this thesis, and provide a rigorous starting point for other researchers exploring task-based language in email communication. In the following chapters, we make direct use of the definitions, analysis and the annotated corpora from the annotation experiments that we have described in this chapter to develop and evaluate automated request and commitment classifiers.

"Most people treat the inbox as an on-line to-do list"
Wendy E. Mackay (1988)

4 Classifying Requests and Commitments at the Message Level

As we have seen in earlier chapters, people are frequently overwhelmed by the number of email messages arriving in their inboxes, many of which contain obligation acts—requests and commitments—that require attention and action. Building on our analysis of requests and commitments of Chapter 3, this chapter presents a computational approach to automatically identifying workplace email messages where people are placing obligations for future action upon themselves and others. Such identification facilitates more efficient triaging of actionable content from the incoming stream of email messages. For now, we focus on identifying obligation acts at the message level (i.e., identifying whether a message does or does not convey a request or a commitment); Chapter 5 examines finer-grained classification (i.e., below the message level).

We approach identifying requests and commitments as a text classification task, making use of supervised, statistical machine learning algorithms to identify messages that convey requests or commitments. As we saw in Chapter 3, the function of conveying an obligation act does not neatly map to a particular set of language forms, making automatically identifying these acts a challenging task.

The data we use to train our classifiers, presented in Section 4.1, is derived from the annotation experiments that we described in Section 3.2. In Section 4.2, we present the features we employ for both request and commitment classification. Section 4.3 then summarises the performance of the request and commitment classifiers in cross-validation experiments over the labelled training data.

Analysing the errors made by these initial classifiers revealed a significant number of errors that arose from the processing of content in parts of a message that were not written by the current sender, such as quoted reply content and legal disclaimers.

Empirically, we found such text was not likely to contain actionable information.[1] Based on this analysis, we hypothesised that identifying and excluding some sections of text from each email message would reduce confounding cues being detected in irrelevant parts of these messages, and thus improve classification performance.

For request and commitment classification, we specifically wish to analyse text within some parts of a message, such as content written by the current sender, and ignore text from other parts, such as advertising content and email signatures. In Section 4.4, we describe Zebra, an automated system we built to segment the body text of email messages into their different functional parts or email zones. We began the design of Zebra through the analysis of a separate collection of workplace email messages, to identify the different functional roles that can be performed by sections of text in an email message. This allowed us to empirically identify nine functional roles that we describe in Section 4.4.3. We then manually annotated zone information across a collection of messages drawn from the Enron email dataset to provide training data for Zebra.

Despite acknowledgement in the literature of the confounding effects of text from different email zones on various classification tasks, the effects of applying email zoning to assist in classification have not previously been explored. In Section 4.5 we demonstrate the performance lift we achieve when identifying request and commitment messages by employing Zebra to filter out text from irrelevant zones prior to classification. We find that using Zebra to focus the obligation act classifiers on only the relevant sections of text in each email message improves both request and commitment classification accuracy, even in the face of errors made during the email zoning task. We also analyse the contribution of lexical features to request and commitment classification, discuss classification learning curves, and present a detailed analysis that explores the sources of classification errors.

Finally, in Section 4.6 we summarise the results of our message-level classification experiments, and the insights we draw from the results to inform our approach to the finer-granularity classification experiments that we present in Chapter 5.

4.1 Training Data for Message-Level Classification

Given an email message as input, complete with header information and body text, the request and commitment classifiers output a binary prediction of whether the message conveys either a request or a commitment. Usually, this equates to finding the presence or absence of one or more request or commitment utterances within the message, however there are cases where an obligation act can be conveyed by a message without any particular utterance performing the act. We saw an example earlier in Figure 3.35(c), which presented an email containing an attached document with no other text in the body. While far from unanimous, more than a quarter of our surveyed respondents interpreted this message as conveying a request.

[1] This content can, however, act as context for interpreting the content written by the current sender.

Both our request and commitment classifiers are built on Support Vector Machine (SVM) classification models, using the implementation available in the Weka machine learning toolkit (Witten and Frank, 2005). SVMs have been previously shown to outperform alternate supervised machine learning methods for text classification in general, and speech act classification in particular — e.g., (Joachims, 1998; Fernandez and Picard, 2002; Khoo, Marom, and Albrecht, 2006). As SVMs are a supervised machine learning method, we require labelled samples to train the models to be able to classify requests and commitments. Training a Support Vector Machine builds a model that classifies unseen data into one class or the other, in our case predicting either the presence or absence of requests/commitments in a message. The SVM model constructs a hyperplane or a set of hyperplanes in a higher-dimensional space and represents the labelled training examples as points in that higher-dimensional space. The example data points are mapped so that examples of different classes (e.g., request and non-request) are maximally far apart. A key insight is that mapping the original finite-dimensional space into a much higher dimensional space can make this separation easier. A good separation between different classes of examples is achieved when the hyperplane that separates examples of each class is maximally distant from the nearest training data points of any class. The model then assigns unseen examples to a predicted class based on which side of the hyperplane they are mapped to.

The training data for both our classifiers is drawn from the labelled data generated from our annotation experiments, as summarised in Section 3.2 and described in more detail in Appendix B. Our request classifier uses the unanimously agreed annotated messages from a set of 664 triply annotated messages as training data. The three-way κ agreement for requests across the complete annotated corpus is 0.681, and after removing messages where annotators disagreed, the training set for the request classifier consists of 505 email messages. We remove messages where annotators disagreed in order to mitigate the effects of annotation noise, as discussed in (Beigman and Klebanov, 2009), where the inclusion of hard cases with unreliable annotations in training data can result in models that make incorrect predictions for uncontroversial cases.

Our commitment classifier draws its training set from a slightly larger collection of 699 email messages that were also annotated by three annotators.[2] The unanimously agreed subset used for training and evaluating our commitment classifier contains 650 messages.

It is interesting to contrast the training datasets used for the request and commitment classifiers. Apart from the difference in size, the annotated data for commitments is much less balanced than that for our request classifier, due to the lower frequency of commitments in email messages; less than 24% of commitment training messages contained a commitment, compared with our 52% of email messages that contain a request in the request training set. The higher percentage of agreed messages in our annotated datasets (93% for commitments *vs.* 78% for requests) largely reflects the fact that more messages in the annotated corpus unambiguously contain no commitment.

[2] The commitment classifier was developed slightly later, when additional annotated data was available. Its training data is drawn from a superset of the same annotated corpus as the request classifier (i.e., the same annotated dataset with more examples annotated by the time we extracted training data for the commitment classifier).

Symbol Used	Pattern
numbers	Any sequence of digits
day	Day names or abbreviations
pronoun-accusative	Accusative pronouns: *me*, *her*, *him*, *us*, *them*
pronoun-nominative	Nominative pronouns: *I*, *she*, *he*, *we*, *they*, *you*
file type	.doc, .pdf, .ppt, .txt, .xls, .rtf
multi-dash	3 or more sequential '–' characters
multi-underscore	3 or more sequential '_' characters

Table 4.1: The normalisation we apply to n-gram features

Despite this agreement on negative examples, our inter-annotator agreement remains significantly lower for commitments than for requests.

4.2 Features for Message-Level Classification

In this section we first discuss the n-gram features used across both classifiers, then focus on the more specific features used for request and commitment classification.

4.2.1 N-gram Features

Both our classifiers rely on lexical features and other cues to identify requests and commitments in each email message. The primary form of these lexical features consists of word n-gram features that are generated from the training data. Before generating word n-gram features, we normalise the message text as shown in Table 4.1, in a manner similar to Carvalho and Cohen (2006). We also add tokens marking the start and end of sentences, detected using a modified version of Scott Piao's sentence splitter (Piao, Wilson, and McEnery, 2002), and tokens marking the start and end of the message body text. This allows our classifiers to capture sentence position information for words in the n-gram feature set. Importantly, our classifiers can then distinguish between a word occurring as the first or last word in a sentence and the same word occurring in other positions within the sentence. This turns out to be useful. For example, a sentence-initial *please*, as in the utterance *Please send me the document*, is a more consistent marker of a request than the same *please* found mid-sentence, as in the utterance *I hope that will please him!*.

When generating our n-gram features, we discard words that, after normalisation, occur less than 3 times in our training data. This cutoff value was identified empirically, and is clearly dependent on the size of our annotated datasets, and reduces the set of unique n-grams from approximately 83,000 to around 73,000 after pruning.

There are some sources of ambiguity that can lead to errors in our feature normalisation and abstraction. A case in point is the word *you*, which can occur as either an accusative or nominative pronoun. For the sake of simplicity, we assume all forms of *you* are nominative. While this may seem questionable from a linguistic perspective, empirical experimentation with more principled classifications or with a strategy that assumes all instances are accusative made little difference to classification performance overall. The former, principled approach was tested by hand-coding instances of *you* as either nominative or accusative features, based on a manual analysis of a test set of data. The differences in performance between using these hand-coded features and using an approach that marked all instances of *you* as either nominative or all as accusative were not statistically significant. This led us to use our simple, if naïve feature encoding for instances of *you*.

4.2.2 Request Features

The specific features we use for our request classifier are:

1. Message length in characters [Integer];
2. Message length in words [Integer];
3. Number of capitalised words [Integer];
4. Percentage of capitalised words [Integer];
5. Number of non alpha-numeric characters [Integer];
6. Percentage of non alpha-numeric characters [Integer];
7. Presence of Subject line reply or forward markers (e.g. `Re:`, `Fw:`) [Boolean];
8. Number of times sender name or initials appear in body text [Integer];
9. Number of times a recipient name or initials appear in body text [Integer];
10. Number of sentences that begin with a modal verb (e.g., *might*, *may*, *should*, *would*) [Integer];
11. Percentage of sentences that begin with a modal verb [Integer];
12. Number of sentences that begin with a question word (e.g, *who*, *what*, *where*, *when*, *why*,
13. Percentage of sentences that begin with a question word [Integer];
14. Number of sentences that end with a question mark [Integer];
15. Percentage of sentences that end with a question mark [Integer]; and
16. Binary word unigram and word bigram features for n-grams that occur at least three times across the training set [Boolean].

These features are motivated by a range of intuitions and observations of requests in email. Sentences that begin with modal verbs or question words or end with a question mark, for example, are frequently associated with interrogative form utterances that often represent requests. Features related to the message length aim to capture any tendency for requests to occur in shorter or longer messages. The presence of recipients' names or initials in the body text can indicate addressivity that may be related to requests being made. Finally the word unigram and bigram features aim to learn lexical patterns that are indicative of requests, such as a sentence-initial *please*.

4.2.3 Commitment Features

The features used for classification of commitments are similar to those used for request classification:

1. Message length in characters [Integer];
2. Message length in words [Integer];
3. Number of capitalised words [Integer];
4. Percentage of capitalised words [Integer];
5. Number of non alpha-numeric characters [Integer];
6. Percentage of non alpha-numeric characters [Integer];
7. Presence of Subject line reply or forward markers (e.g., `Re:`, `Fw:`) [Boolean];
8. Number of times sender name or initials appear in body text [Integer];
9. Number of times a recipient name or initials appear in body text [Integer]; and
10. Binary word unigram and word bigram features for n-grams that occur at least three times across the training set [Boolean].

As for requests, these features are motivated by a range of intuitions and observations. Messages that contain reply markers in the subject line may contain a commitment in response to an initial request. Features related to the message length aim to capture any tendency for commitments to occur in shorter or longer messages. The presence of the sender's name or initials in the body text can indicate a commitment expressed in the third-person, sometimes seen as a list of action items assigned to specific people. As for requests, the word unigram and bigram features aim to learn lexical patterns that are indicative of commitments, such as *I will* or *I promise*.

4.3 Message-Level Obligation Act Classification Performance

We created our initial classifiers using Weka's SVM implementation, using the training data described in Section 4.1. Because gathering labelled email data is an expensive process, in terms of the time and effort involved, we do not have a held-out testing set of labelled email messages. Instead, we use stratified 10-fold cross validation, a widely accepted technique for repeatedly sampling training and test data from the same underlying labelled dataset, and report the average performance of the classifier across these runs.

4.3.1 Request Classification Results

Our initial request classifier achieved a classification accuracy of 72.28%. Table 4.2 shows accuracy, precision, recall and F-score results, calculated using stratified 10-fold cross validation, compared against a majority class baseline. Given the well-balanced

	Majority Baseline		Classifier without Zoning	
	Request	Non-Request	Request	Non-Request
Accuracy	52.08%		72.28%	
Precision	0.521	0.000	0.729	0.716
Recall	1.000	0.000	0.745	0.698
F-score	0.685	0.000	0.737	0.707

Table 4.2: Our message-level request classifier results.

nature of our training data (52.08% of messages contain a request), this is a reasonable basis for comparison.

We performed a variety of feature-ablation experiments on our initial request classifier to determine the most salient features, and to examine the contribution of various feature classes. As expected, lexical information was crucial to accurate request classification. We first experimented with removing all lexical (n-gram) features. When doing so, our message-level request classifier accuracy dropped from 72.28% to 57.62%. In contrast, when we apply only n-gram features, removing all other features, we still achieve an accuracy of 71.49%. This emphasises the importance of lexical features for this initial classifier; the remaining features together boost accuracy by less than 1%.

Using an Information Gain metric, we ranked the n-gram features in terms of their usefulness. Table 4.3 shows the top-10 unigrams and bigrams for the request classifier. Using these top-10 n-grams (plus our non-n-gram features), we achieve only 66.34% accuracy. These top-10 n-grams include words and phrases such as *parsing*, *iso final* and *westdesk* /that do not seem to align well with linguistic intuitions about the lexical items we might expect to see in requests. When we examined these results further, we discovered several similar, apparently automated, messages that were annotated (as non-requests) in our training corpus which appear to be the source of several of these unexpected top-10 n-grams. An example of one of these messages is shown in Figure 4.1. That these messages are the source of a number of the highly ranked n-gram features strongly suggests that the classifier is not learning features from the training set at a useful level of generality.

One approach might be to attempt to filter out such messages before looking for requests and commitments. This approach does, however, bring with it a significant set of new challenges in terms of how to identify messages that should be filtered and ignored. As we outlined in Section 4.4, we pursue an alternate approach that is more widely applicable and that addresses the presence of such messages and other confounding phenomena in a more flexible manner.

We also analysed the prediction errors from this initial request classifier, and uncovered a series of classification errors that appeared to be due to request-like signals being picked up from irrelevant sections of email text, such as email signatures and

Word Unigrams	Word Bigrams	
	Word 1	Word 2
PRONOUN-ACCUSATIVE	let	PRONOUN-ACCUSATIVE
please	PRONOUN-ACCUSATIVE	know
iso	START-SENTENCE	no
PRONOUN-ACCUSATIVE	start	date
hourahead	hour	:
attached	;	hourahead
let	hourahead	hour
westdesk	START-SENTENCE	start
parsing	westdesk	/
if	iso	final

Table 4.3: The top 10 useful n-grams for our request classifier without zoning, ranked by Information Gain.

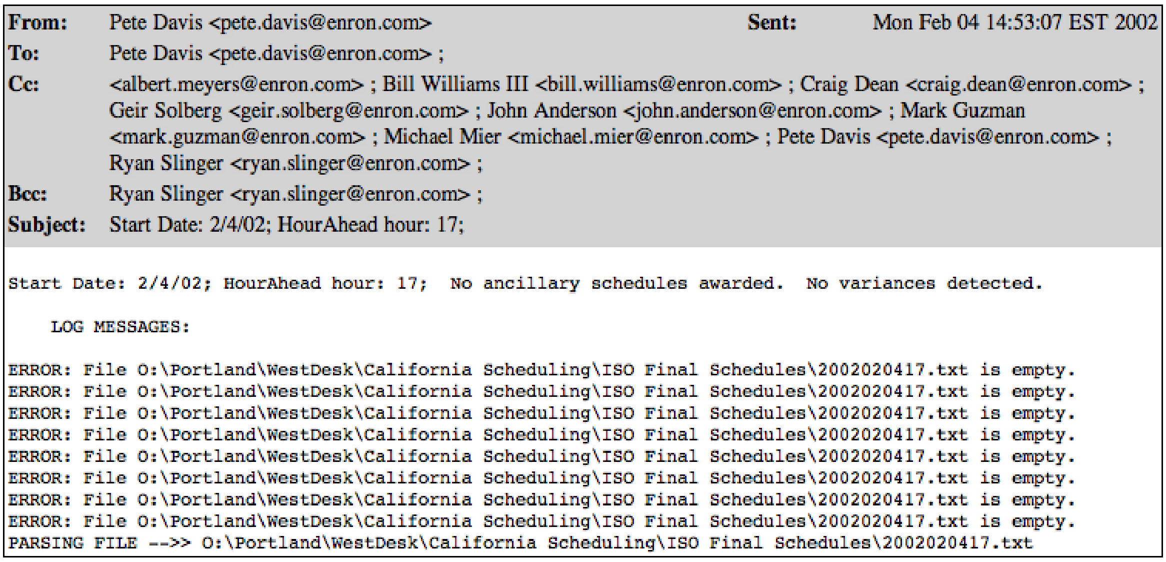
quoted reply content. An example is shown below in Figure 4.2. In this example, the highlighted utterance in the footer of the email, *Do You Yahoo!?*, looks like an interrogative request. This section of content, however, has been automatically appended to the message rather than deliberately added by the sender. Despite having the surface form of a request, this utterance functions as advertising. If we were able to identify the function of this section of text, our request classifier would never attempt to identify obligation acts within this text.

Together with the feature analysis described above, this observation led us to believe that our request classifier would benefit from a mechanism that could identify and ignore some message parts. We describe our approach to identifying relevant and irrelevant message parts in Section 4.4.

4.3.2 Commitment Classification Results

Our initial commitment classifier achieved classification accuracy of 82.76% against a baseline majority class accuracy of 76.31%. These results are calculated over our corpus of 650 unanimously annotated email messages. Table 4.4 shows the accuracy, precision, recall and F-score results, compared against our majority class baseline results. Again, all metrics are calculated using stratified 10-fold cross validation.

As we noted earlier, the data for our commitment classifier is more skewed than that for our request classifier, due to only 24% of email messages conveying a commitment. This imbalance leads to much better performance by the majority baseline system, as it can achieve 76% accuracy simply by always predicting that messages do not contain a commitment. Our initial commitment classifier does better than this baseline, as shown in Table 4.4.



From: Pete Davis <pete.davis@enron.com> **Sent:** Mon Feb 04 14:53:07 EST 2002
To: Pete Davis <pete.davis@enron.com> ;
Cc: <albert.meyers@enron.com> ; Bill Williams III <bill.williams@enron.com> ; Craig Dean <craig.dean@enron.com> ; Geir Solberg <geir.solberg@enron.com> ; John Anderson <john.anderson@enron.com> ; Mark Guzman <mark.guzman@enron.com> ; Michael Mier <michael.mier@enron.com> ; Pete Davis <pete.davis@enron.com> ; Ryan Slinger <ryan.slinger@enron.com> ;
Bcc: Ryan Slinger <ryan.slinger@enron.com> ;
Subject: Start Date: 2/4/02; HourAhead hour: 17;

```
Start Date: 2/4/02; HourAhead hour: 17;  No ancillary schedules awarded.  No variances detected.

    LOG MESSAGES:

ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
ERROR: File O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt is empty.
PARSING FILE -->> O:\Portland\WestDesk\California Scheduling\ISO Final Schedules\2002020417.txt
```

Figure 4.1: One example of a collection of similar email messages which were marked as non-requests that were the source of several highly-ranked n-grams for our request classifier without email zoning.

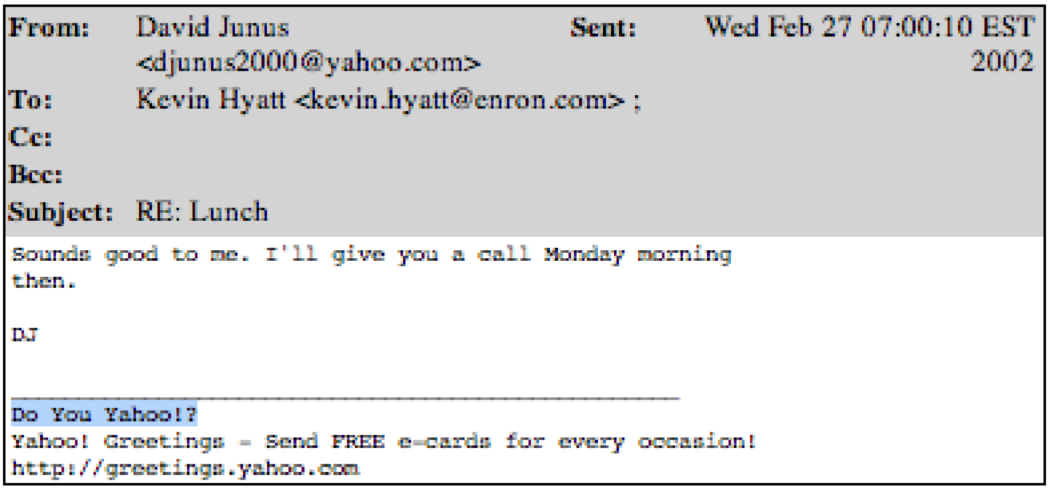
	Majority Baseline		All Features	
	Commit	Non-Commit	Commit	Non-Commit
Accuracy	76.31%		82.77%	
Precision	0.000	0.763	0.809	0.830
Recall	0.000	1.000	0.357	0.978
F-score	0.000	0.866	0.495	0.896

Table 4.4: Our message-level commitment classifier results.

Unlike the request classifier, where recall and precision for both request and non-request messages was approximately equal, we see much more variation in the commitment classification results. For commitment messages, recall in particular is much poorer than for non-commitment messages. This is likely to be due to the weaker lexical cues that exist for commitments. As discussed in Chapter 3, the surface form of a commitment is frequently indistinguishable from a prediction or simple statement about the future that conveys no obligation. For example, distinguishing between the prediction in Example (4.1) and the commitment in Example (4.2) is difficult to do based on surface features alone.

(4.1) *Our luck will change tomorrow.*

(4.2) *Our meeting will be tomorrow.*



From: David Junus <djunus2000@yahoo.com>
Sent: Wed Feb 27 07:00:10 EST 2002
To: Kevin Hyatt <kevin.hyatt@enron.com> ;
Cc:
Bcc:
Subject: RE: Lunch

```
Sounds good to me. I'll give you a call Monday morning
then.

DJ

__________________________________________________
Do You Yahoo!?
Yahoo! Greetings - Send FREE e-cards for every occasion!
http://greetings.yahoo.com
```

Figure 4.2: An email showing a highlighted interrogative utterance in the signature zone of an email.

As the majority of our features are lexically based, the performance of our classifier is therefore affected by such ambiguity. In contrast, there are stronger and more reliable lexical cues for recognising many classes of requests.

Again, we used Information Gain to rank the n-gram features in terms of their usefulness for commitment classification. Table 4.5 shows the top-10 unigrams and bigrams for our non-zoning commitment classifier. Using these top-10 n-grams (plus our non-n-gram features), we achieve only 78.30% accuracy, which is still 2% above our baseline accuracy.

We also analysed the prediction errors from our initial commitment classifier, and while the influence did not seem as significant as for requests, we did uncover several classification errors that appeared to be due to commitment-like signals being picked up from parts of messages such as email signatures and quoted reply content. An example is shown in Figure 4.3, where a commitment-like utterance was identified in the signature of an email. We confirmed that it was this utterance that triggered the commitment classifier by removing it, and reclassifying the message, which resulted in a negative commitment output.

These observations reinforced the formulation of our hypothesis that our obligation act classifiers would benefit from an automated classifier that could identify and ignore such message parts. In the next section, we present our email zone classifier that identifies the different functional sections of text within an email message.

Word Unigrams	Word Bigrams	
	Word 1	Word 2
PRONOUN-NOMINATIVE	PRONOUN-NOMINATIVE	will
will	would	like
be	will	be
let	let	PRONOUN-ACCUSATIVE
would	like	to
discuss	PRONOUN-NOMINATIVE	would
if	the	week
click	PRONOUN-ACCUSATIVE	know
the	PRONOUN-NOMINATIVE	am
attending	to	discuss

TABLE 4.5: The top 10 useful n-grams for our commitment classifier without zoning, ranked by Information Gain.

4.4 Email Zoning

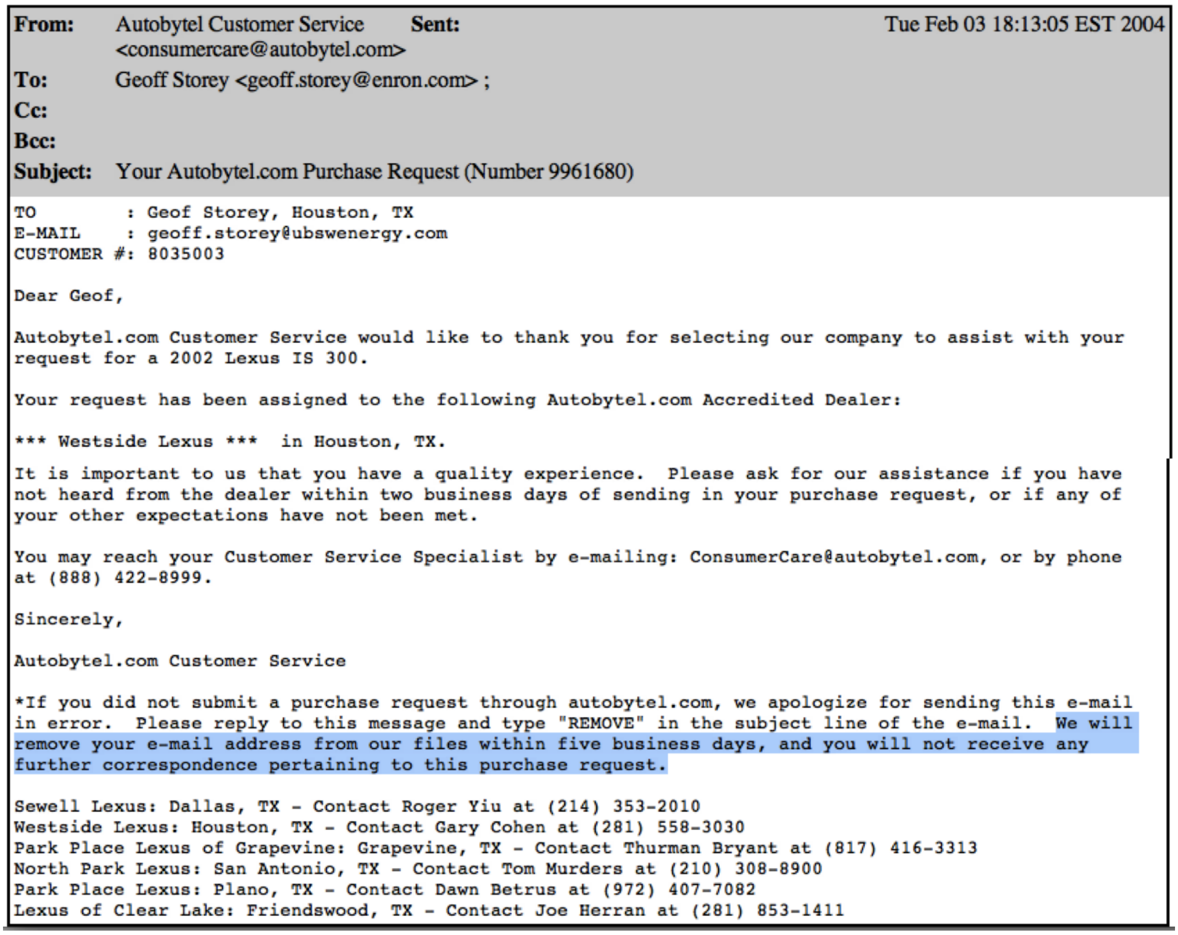
Although not marked in any consistent or machine-readable manner, email message bodies consist of different functional parts such as email signatures, quoted reply content and advertising content. We refer to each different functional section of text within the body text of an email message as an EMAIL ZONE, and the task of identifying these zones within a message as EMAIL ZONING. Common zones within an email message include email signatures, advertising, legal disclaimers and quoted reply text.

The task of identifying email zones takes the homogeneous block of body text from an email message, like that shown in Figure 4.4, and automatically identifies the structure shown in Figure 4.5.

Many language processing tools stand to benefit from better knowledge of this message structure, facilitating focus on relevant content in specific parts of a message. In particular, access to zone information would allow email classification, summarisation and analysis tools to separate or filter out 'noise' and focus on the content in specific zones of a message that are relevant to the application at hand.

Email contact mining tools such as that reported in (Culotta, Bekkerman, and McCallum, 2004), for example, might access the email signature zone, while tools that attempt to identify tasks or action items in email, as developed, for example, in (Bellotti et al., 2003; Corston-Oliver et al., 2004; Bennett and Carbonell, 2007; Lampert, Paris, and Dale, 2007), might restrict themselves to the sender-authored and forwarded content. Despite previous work on this problem, there are no available tools that can reliably extract or identify the different functional zones of an email message.

While there is no agreed standard set of email zones, there are clearly different functional parts within the body text of email messages. For example, the content of an email disclaimer is functionally different from the sender-authored content and from the quoted reply content automatically included from previous messages in the thread



From: Autobytel Customer Service <consumercare@autobytel.com> **Sent:** Tue Feb 03 18:13:05 EST 2004
To: Geoff Storey <geoff.storey@enron.com> ;
Cc:
Bcc:
Subject: Your Autobytel.com Purchase Request (Number 9961680)

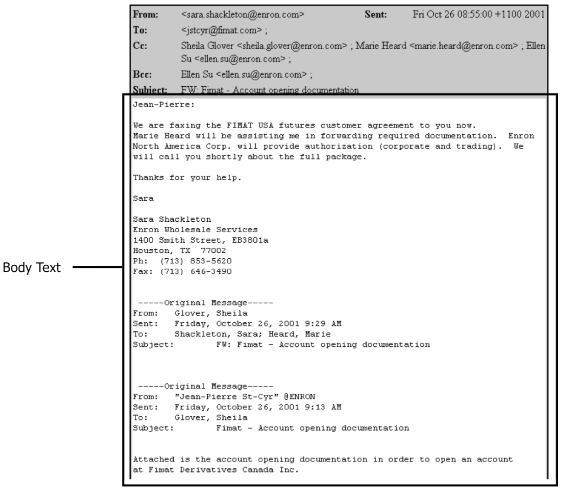
```
TO          : Geof Storey, Houston, TX
E-MAIL      : geoff.storey@ubswenergy.com
CUSTOMER #: 8035003

Dear Geof,

Autobytel.com Customer Service would like to thank you for selecting our company to assist with your
request for a 2002 Lexus IS 300.

Your request has been assigned to the following Autobytel.com Accredited Dealer:

*** Westside Lexus ***  in Houston, TX.

It is important to us that you have a quality experience.  Please ask for our assistance if you have
not heard from the dealer within two business days of sending in your purchase request, or if any of
your other expectations have not been met.

You may reach your Customer Service Specialist by e-mailing: ConsumerCare@autobytel.com, or by phone
at (888) 422-8999.

Sincerely,

Autobytel.com Customer Service

*If you did not submit a purchase request through autobytel.com, we apologize for sending this e-mail
in error.  Please reply to this message and type "REMOVE" in the subject line of the e-mail.  We will
remove your e-mail address from our files within five business days, and you will not receive any
further correspondence pertaining to this purchase request.

Sewell Lexus: Dallas, TX - Contact Roger Yiu at (214) 353-2010
Westside Lexus: Houston, TX - Contact Gary Cohen at (281) 558-3030
Park Place Lexus of Grapevine: Grapevine, TX - Contact Thurman Bryant at (817) 416-3313
North Park Lexus: San Antonio, TX - Contact Tom Murders at (210) 308-8900
Park Place Lexus: Plano, TX - Contact Dawn Betrus at (972) 407-7082
Lexus of Clear Lake: Friendswood, TX - Contact Joe Herran at (281) 853-1411
```

Figure 4.3: An email message with a highlighted phrase that looks like a commitment and was identified by our commitment classifier, but occurs in the email signature, and was not marked by human annotators.

of conversation. Of course, there are different distinctions that can be drawn between zones; in this section we explore several different categorisations based on a proposed set of nine underlying email zones.

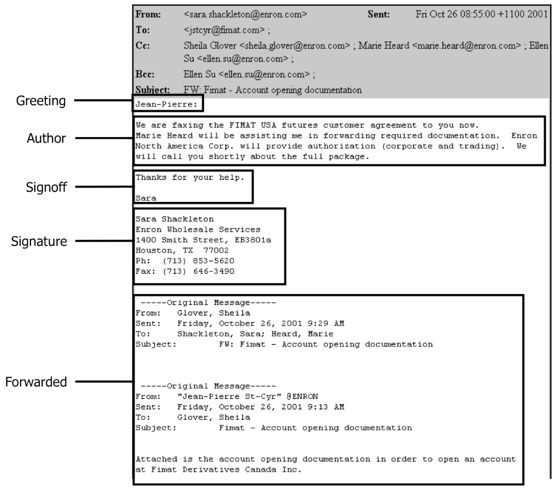
Segmenting email messages into zones is a challenging task. Accurate segmentation is hampered by the lack of standard syntax used by different email clients to indicate different message parts, and by the *ad hoc* ways in which people vary the structure and layout of their messages. When replying to a message, for example, it is often useful to include all or part of the original message that is being replied to. Different email clients indicate quoted material in different ways. By default, some prefix every line of the quoted message with a character such as '>' or '|', while others indent the quoted content or insert the quoted message unmodified, prefixed by a message header.

Sometimes the new content is above the quoted content, as in Figure 3.24 (a style known as TOP-POSTING); in other cases, the new content may appear after the quoted



```
From:     <sara.shackleton@enron.com>          Sent:     Fri Oct 26 08:55:00 +1100 2001
To:       <jstcyr@fimat.com> ;
Cc:       Sheila Glover <sheila.glover@enron.com> ; Marie Heard <marie.heard@enron.com> ; Ellen
          Su <ellen.su@enron.com> ;
Bcc:      Ellen Su <ellen.su@enron.com> ;
Subject:  FW: Fimat - Account opening documentation
Jean-Pierre:

We are faxing the FIMAT USA futures customer agreement to you now.
Marie Heard will be assisting me in forwarding required documentation.  Enron
North America Corp. will provide authorization (corporate and trading).  We
will call you shortly about the full package.

Thanks for your help.

Sara

Sara Shackleton
Enron Wholesale Services
1400 Smith Street, EB3801a
Houston, TX  77002
Ph:  (713) 853-5620
Fax: (713) 646-3490


 -----Original Message-----
From:   Glover, Sheila
Sent:   Friday, October 26, 2001 9:29 AM
To:     Shackleton, Sara; Heard, Marie
Subject:        FW: Fimat - Account opening documentation



 -----Original Message-----
From:   "Jean-Pierre St-Cyr" @ENRON
Sent:   Friday, October 26, 2001 9:13 AM
To:     Glover, Sheila
Subject:        Fimat - Account opening documentation


Attached is the account opening documentation in order to open an account
at Fimat Derivatives Canada Inc.
```

Figure 4.4: A standard email message with unstructured body text.



```
From:     <sara.shackleton@enron.com>          Sent:     Fri Oct 26 08:55:00 +1100 2001
To:       <jstcyr@fimat.com> ;
Cc:       Sheila Glover <sheila.glover@enron.com> ; Marie Heard <marie.heard@enron.com> ; Ellen
          Su <ellen.su@enron.com> ;
Bcc:      Ellen Su <ellen.su@enron.com> ;
Subject:  FW: Fimat - Account opening documentation
Jean-Pierre:

We are faxing the FIMAT USA futures customer agreement to you now.
Marie Heard will be assisting me in forwarding required documentation.  Enron
North America Corp. will provide authorization (corporate and trading).  We
will call you shortly about the full package.

Thanks for your help.

Sara

Sara Shackleton
Enron Wholesale Services
1400 Smith Street, EB3801a
Houston, TX  77002
Ph:  (713) 853-5620
Fax: (713) 646-3490

 -----Original Message-----
From:   Glover, Sheila
Sent:   Friday, October 26, 2001 9:29 AM
To:     Shackleton, Sara; Heard, Marie
Subject:        FW: Fimat - Account opening documentation



 -----Original Message-----
From:   "Jean-Pierre St-Cyr" @ENRON
Sent:   Friday, October 26, 2001 9:13 AM
To:     Glover, Sheila
Subject:        Fimat - Account opening documentation


Attached is the account opening documentation in order to open an account
at Fimat Derivatives Canada Inc.
```

Figure 4.5: The same email message with email zones identified.

content (bottom-posting) or interleaved with the quoted content (inline replying). Confounding the issue further is that users are able to configure their email client to suit their individual tastes, and can change both the syntax of quoting and their quoting style (top, bottom or inline replying) on a per-message basis.

To address these challenges, in this section we describe Zebra, our email zone classification system. We start by reviewing related work in Section 4.4.1. Sections 4.4.2 and 4.4.3 then present details of the email data we use for training our zone classifier. In Section 4.4.4, we describe two complimentary approaches to zone classification, one that is line-based and one that is fragment-based. Our classification features are presented in Section 4.4.5, followed by an evaluation of Zebra's performance across two, three and nine email zone classification tasks in Section 4.4.6.

4.4.1 Related Work

Automating the zoning of emails into their functional parts involves both text segmentation and classification. The main focus of most work on text segmentation is topic-based segmentation of news text, as in the work of (Hearst, 1997; Beeferman, Berger, and Lafferty, 1997), but there has also been recent work in applying both supervised and unsupervised topic segmentation to email text (Joty et al., 2010; Joty et al., 2011). This work has found that incorporating conversational structure into the topic segmentation models produces more accurate topic segmentations.

Aside from topic segmentation, there have also been some previous attempts at identifying functional zones in email messages. Chen, Hu, and Sproat (1999) looked at both linguistic and two-dimensional layout cues for extracting structured content from email signature zones in email messages. The focus of their work was on extracting information from already identified signature blocks using a combination of two-dimensional structural analysis and one-dimensional grammatical constraints; the intended application domain was as a component in a system for email text-to-speech rendering. The authors claim that their system can be modified to also identify signature blocks within email messages, but their system performs this task with a recall of only 53%. No attempt is made to identify functional zones other than email signatures.

Carvalho and Cohen's (Carvalho and Cohen, 2004) Jangada system attempted to identify email signatures within plain text email messages and to extract email signatures and reply lines. Unfortunately, the 20 Newsgroups corpus[3] they worked with contains 15-year-old Usenet messages which are much more homogeneous in their syntax than contemporary email, particularly in terms of how quoted text from previous messages is indicated. As a result, using a very simple metric (a line-initial '>' character) to identify reply lines achieves more than 95% accuracy. In contrast, this same simple metric applied to the Enron email data we annotated detects less than 10% of actual reply or forward lines.

Usenet messages are also markedly different from contemporary email when it comes to email signatures. Most Usenet clients produced messages which conformed to RFC3676 (Gellens, 2004), a standard that formalised a "long-standing convention in Usenet news ... of using two hyphens -- as the separator line between the body and the signature of a message." Unfortunately, this convention has long since ceased to be observed in email messages. Carvalho and Cohen's email signature detection approach also benefits greatly from a simplifying assumption that signatures are found in the last 10 lines of an email message. While this holds true for their Usenet message data, it is no longer the case for contemporary email.

In attempting to use Carvalho and Cohen's system to identify signature blocks and reply lines in our own work, we identified similar shortcomings to those noted by Estival et al. (2007). In particular, Jangada did not accurately identify forwarded or reply content in email data from the Enron email corpus. We believe that the use of older Usenet-style messages to train Jangada is a significant factor in the systematic errors the system makes in failing to identify quoted reply, forwarded and signature content in messages formatted in the range of message formats and styles popularised by Microsoft Outlook. These errors are a fundamental problem with Jangada, especially since Outlook is the most common client used to compose messages in our annotated email collection drawn from the Enron corpus. More generally, we note that Outlook is the most popular email client in current use, with an estimated 350–400 million users worldwide,[4] representing anywhere up to 40% of all email users.[5]

More recently, as part of their work on profiling authors of email messages, Estival et al. (2007) classified email bodies into five email zones. Their paper does not provide results for five-zone classification, but they report accuracy of 88.16% using a Conditional Random Field (CRF) classifier to distinguish three zones: reply, author

[3] Available from http://people.csail.mit.edu/jrennie/20Newsgroups/.

[4] Xobni Co-founder Adam Smith and former Engineering VP Gabor Cselle have both published Outlook user statistics. See http://www.xobni.com/asmith/archives/66 and http://gaborcselle.com/blog/2008/05/xobnis-journey-to-right-product.html.

[5] Campaign Monitor regularly publishes statistics on email software usage at http://www.campaignmonitor.com/stats/email-clients/.

and signature. We use their classification scheme as the starting point for our own set of email zones that we present here.

4.4.2 Email Zoning Data and Annotation

The training data for our zone classifier consists of 11881 annotated lines from almost 400 email messages drawn at random from the Enron email corpus that we use for our request and commitment annotation experiments. Following Estival et al. (2007), we used only a single annotator since the task revealed itself to be relatively uncontroversial. Each line in the body text of selected messages was marked by the annotator (one of the authors) as belonging to one of nine zones. After removing blank lines, which we do not attempt to classify, we are left with 7922 annotated lines as training data for Zebra. The frequency of each zone within this annotated dataset is shown in Table 4.8.

Annotation was performed using a modified version of the custom web application used to annotate requests and commitments. This application displays email messages using a look-and-feel that approximates the presentation used in Microsoft Outlook. The standard header fields and values (From, Date, To, Cc, Bcc, and Subject) are shown preceding the message content, which is presented as a sequence of lines. For each line, annotators select a single annotation value from a drop-down menu.

4.4.3 Email Zones

As noted earlier, we refer to the different functional components of email messages as EMAIL ZONES. The zones we propose refine and extend the five categories — *Author Text*, *Signature*, *Advertisement* (automatically appended advertising), *Quoted Text* (extended quotations such as song lyrics or poems), and *Reply Lines* (including forwarded and reply text) — identified by Estival et al. (2007).

We consider that each line of text in the body of an email message belongs to one of nine more fine-grained email zones, which we describe in this section. We intend these nine email zones to be abstracted and adapted to suit different tasks. To illustrate, we present the zones below abstracted into three classes: sender-authored content, boilerplate content, and content quoted from other conversations. This is the zone partition we use to generate the three-zone results reported in Section 4.4.6. This categorisation is useful for problems such as finding action items in email messages: such detection tools would look in text from the sender-authored message zones for new action item information, and could also look in quoted conversation content to link new action item information (such as reported completions) to previous action item content.

4.4.3.1 Sender Zones

Sender zones contain text written by the current email sender. The GREETING and SIGNOFF zones are sub-zones of the *Author* zone, usually appearing as the first and last items respectively in the *Author* zone. Thus, our proposed sender zones are:

1. **Author:** New content from the current email sender. This specifically excludes any text authored by the sender that is included from previous messages.
2. **Greeting:** Terms of address and recipient names at the beginning of a message (e.g., *Dear/Hi/Hey Noam*).
3. **Signoff:** The message closing (e.g., *Thanks/Cheers/Regards, John*).

4.4.3.2 Quoted Conversation Zones

Quoted conversation zones include both content quoted in reply to previous messages in the same conversation thread and forwarded content from other conversations.[6] Our quoted conversation zones are:

4. **Reply:** Content quoted from a previous message in the same conversation thread, including any embedded signatures, attachments, advertising, disclaimers, author content and forwarded content. Content in a reply content zone may include previously sent content authored by the current sender.
5. **Forward:** Content from an email message outside the current conversation thread that has been forwarded by the current email sender, including any embedded signatures, attachments, advertising, disclaimers, author content and reply content.

4.4.3.3 Boilerplate Zones

Boilerplate zones contain content that is reused without modification across multiple email messages. Our proposed boilerplate zones are:

6. **Signature:** Content containing contact or other information that is automatically inserted in a message. In contrast to disclaimer or advertising content, signature content is usually templated content written once by the email author, and automatically or semi-automatically included in email messages. A user may also use a *Signature* in place of a *Signoff*; in such cases, we still mark the text as a *Signature*.
7. **Advertising:** Advertising material in an email message. Such material often appears at the end of a message (e.g., *Do you Yahoo!?*), but may also appear prefixed or inline with the content of the message, (e.g., in sponsored mailing lists).
8. **Disclaimer:** Legal disclaimers and privacy statements, often automatically appended.

[6] Although we recognise the need for the *Quoted Text* zone proposed by Estival et al. (2007), no such data occurs in our collection of annotated email messages. We therefore omit this zone from our current set.

9. **Attachment:** Automated text indicating or referring to attached documents, such as that shown in line 16 of Figure 4.6. Note that this zone does not apply to manually authored reference to attachments, nor to the actual content of attachments (which we do not classify).

4.4.3.4 Other Zone Abstractions

Our nine email zones can also be reduced to a binary scheme to distinguish text authored by the sender from text authored by others. This distinction is useful for problems such as author attribution or profiling tasks. In this two-class case, the sender-authored zones would be *Author*, *Greeting*, *Signoff* and *Signature*, while the other-authored zones would be *Reply*, *Forward*, *Disclaimer*, *Advertising* and *Attachment*. Note that in this binary partition, text which is automatically inserted by an email client or other system, such as advertising content and legal disclaimers, is considered to be text authored by others. This is the partition of zones we use in our two-zone experiments reported in Section 4.4.6.

Figure 4.6 shows an email message with each line marked with the appropriate email zone. Zones are marked according to both the nine fine-grained zones and the the abstracted three-zone scheme described above. Note, however, that not all of the nine fine-grained zones, nor all of the three abstracted zones, are actually present in this particular message. Finally, Figure 4.6 also shows the binary zone scheme, with lines 1–11, bounded by the heavy black box, representing sender-authored text, and the text in lines 13–35 representing text authored by others.

4.4.4 Zone Segmentation and Classification Methods

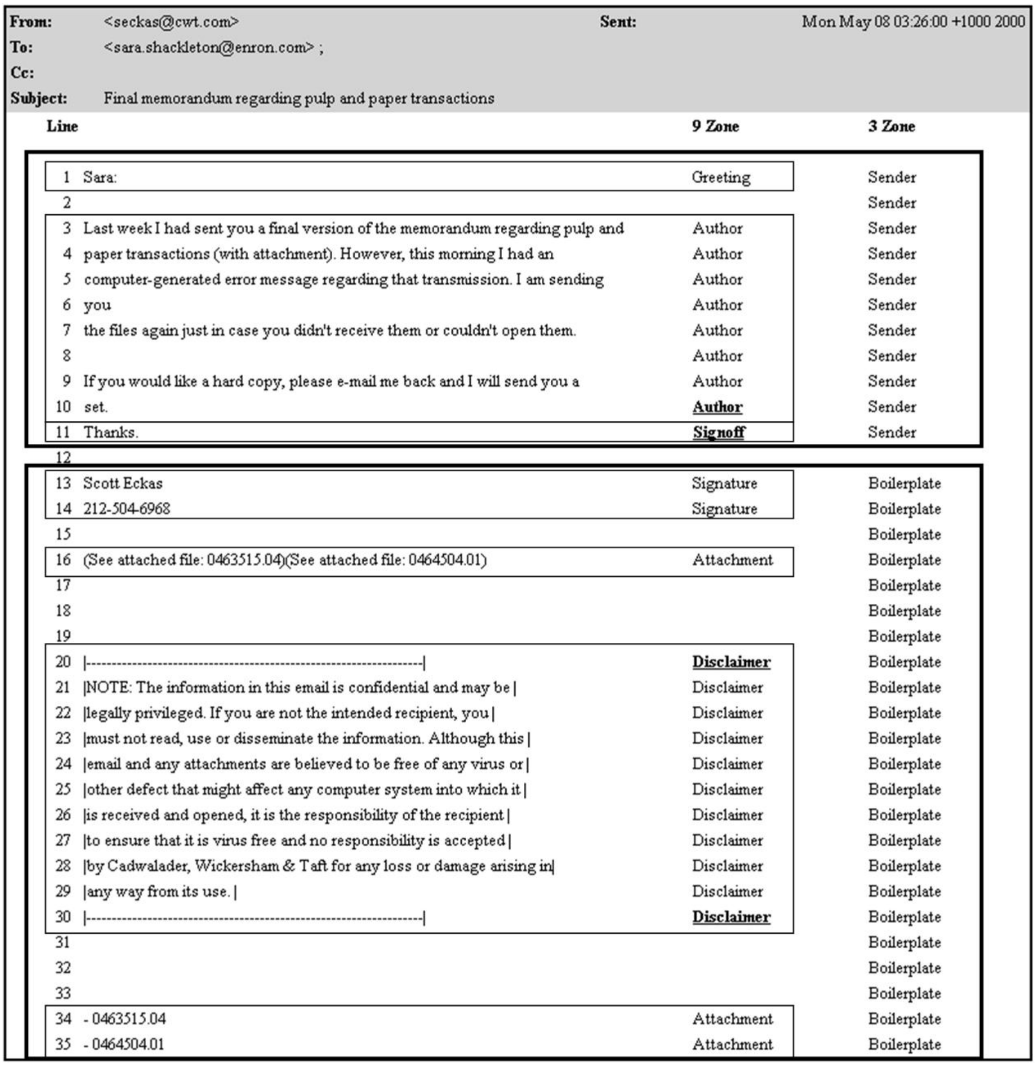
Our email zone classification system is based around an SVM classifier using features that capture graphic, orthographic and lexical information about the content of an email message.

To classify the zones in an email message, we experimented with two approaches. The first employs a two-stage approach that segments a message into zone fragments and then classifies those fragments. Our second method simply classifies lines independently, returning a classification for each non-blank line in an email message. Our hypothesis was that classifying larger text fragments would lead to better performance due to the text fragments containing more cues about the zone type.

4.4.4.1 Fragment-based Zone Segmentation and Classification

Zone fragment classification is a two-step process. First it predicts the zone boundaries using a simple heuristic, then it classifies the resulting ZONE FRAGMENTS, the sets of content lines that lie between these hypothesised boundaries.

In order to determine how well we can detect zone boundaries, we first need to establish the correct zone boundaries in our collection of zone-annotated email messages.



From: <seckas@cwt.com> **Sent:** Mon May 08 03:26:00 +1000 2000
To: <sara.shackleton@enron.com> ;
Cc:
Subject: Final memorandum regarding pulp and paper transactions

<table>
<tr><th>Line</th><th></th><th>9 Zone</th><th>3 Zone</th></tr>
<tr><td>1</td><td>Sara:</td><td>Greeting</td><td>Sender</td></tr>
<tr><td>2</td><td></td><td></td><td>Sender</td></tr>
<tr><td>3</td><td>Last week I had sent you a final version of the memorandum regarding pulp and</td><td>Author</td><td>Sender</td></tr>
<tr><td>4</td><td>paper transactions (with attachment). However, this morning I had an</td><td>Author</td><td>Sender</td></tr>
<tr><td>5</td><td>computer-generated error message regarding that transmission. I am sending</td><td>Author</td><td>Sender</td></tr>
<tr><td>6</td><td>you</td><td>Author</td><td>Sender</td></tr>
<tr><td>7</td><td>the files again just in case you didn't receive them or couldn't open them.</td><td>Author</td><td>Sender</td></tr>
<tr><td>8</td><td></td><td>Author</td><td>Sender</td></tr>
<tr><td>9</td><td>If you would like a hard copy, please e-mail me back and I will send you a</td><td>Author</td><td>Sender</td></tr>
<tr><td>10</td><td>set.</td><td><b><u>Author</u></b></td><td>Sender</td></tr>
<tr><td>11</td><td>Thanks.</td><td><b><u>Signoff</u></b></td><td>Sender</td></tr>
<tr><td>12</td><td></td><td></td><td></td></tr>
<tr><td>13</td><td>Scott Eckas</td><td>Signature</td><td>Boilerplate</td></tr>
<tr><td>14</td><td>212-504-6968</td><td>Signature</td><td>Boilerplate</td></tr>
<tr><td>15</td><td></td><td></td><td>Boilerplate</td></tr>
<tr><td>16</td><td>(See attached file: 0463515.04)(See attached file: 0464504.01)</td><td>Attachment</td><td>Boilerplate</td></tr>
<tr><td>17</td><td></td><td></td><td>Boilerplate</td></tr>
<tr><td>18</td><td></td><td></td><td>Boilerplate</td></tr>
<tr><td>19</td><td></td><td></td><td>Boilerplate</td></tr>
<tr><td>20</td><td>|---|</td><td><b><u>Disclaimer</u></b></td><td>Boilerplate</td></tr>
<tr><td>21</td><td>|NOTE: The information in this email is confidential and may be |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>22</td><td>|legally privileged. If you are not the intended recipient, you |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>23</td><td>|must not read, use or disseminate the information. Although this |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>24</td><td>|email and any attachments are believed to be free of any virus or |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>25</td><td>|other defect that might affect any computer system into which it |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>26</td><td>|is received and opened, it is the responsibility of the recipient |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>27</td><td>|to ensure that it is virus free and no responsibility is accepted |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>28</td><td>|by Cadwalader, Wickersham &amp; Taft for any loss or damage arising in|</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>29</td><td>|any way from its use. |</td><td>Disclaimer</td><td>Boilerplate</td></tr>
<tr><td>30</td><td>|---|</td><td><b><u>Disclaimer</u></b></td><td>Boilerplate</td></tr>
<tr><td>31</td><td></td><td></td><td>Boilerplate</td></tr>
<tr><td>32</td><td></td><td></td><td>Boilerplate</td></tr>
<tr><td>33</td><td></td><td></td><td>Boilerplate</td></tr>
<tr><td>34</td><td>- 0463515.04</td><td>Attachment</td><td>Boilerplate</td></tr>
<tr><td>35</td><td>- 0464504.01</td><td>Attachment</td><td>Boilerplate</td></tr>
</table>

Figure 4.6: An example email message marked with zone annotations. The heavy-edged boxes show the two different zones under a three-zone annotation scheme, while the the lighter-edged boxes show the seven different zones identified using our nine-zone scheme.

Zone Boundaries A zone boundary is defined as a continuous collection of one or more lines that separate two different email zones. Lines that separate two zones and are blank, contain only whitespace or contain only punctuation characters are called buffer lines.

Since classification of blank lines between zones is often ambiguous, empty or whitespace-only buffer lines are not included as content in any zone, and thus are not classified. Instead, they are treated as strictly part of the zone boundary. In Figure 4.6, these lines are shown without any zone annotation. Zone boundary lines that are included as content in a zone have their zone annotation styled in bold and underlined. The important point here is that zone boundaries are specific to a zone classification scheme. For nine-zone classification of the message in Figure 4.6, there are six zone boundaries: line 2, lines 10–11, line 12, line 15, lines 17–20, and lines 30–33. For three-zone classification, the only zone boundary consists of line 12, separating the sender and boilerplate zones.

Based on these definitions, there are three different types of zone boundaries:

1. **Blank boundaries** contain only empty or whitespace-only buffer lines. Lines in these zone boundaries are strictly separate from the zone content. An example is Line 12 in Figure 4.6, for both the three- and nine-zone classification.
2. **Separator boundaries** contain only buffer lines, but must contain at least one punctuation-character buffer line that is retained as content in one or both zones. In Figure 4.6, an example is the zone boundary containing lines 17–20 that separates the *Attachment* and *Disclaimer* zones for nine-zone classification, since line 20 is retained as part of the *Disclaimer* zone content.
3. **Adjoining boundaries** consist of the last content line of the earlier zone and the first content line of the following zone. These boundaries occur where no buffer lines exist between the two zones. An example is the zone boundary containing lines 10 and 11 that separates the *Author* and *Signoff* zones in Figure 4.6 for nine-zone classification.

Hypothesising Zone Boundaries To identify zone boundaries in unannotated email data, we employ a very simple heuristic approach. Specifically, we consider every line in the body of an email message that matches any of the following criteria to be a zone boundary:

1. A blank line;
2. A line containing only whitespace; or
3. A line beginning with four or more repeated punctuation characters, optionally prefixed by whitespace.

Our efforts to apply more sophisticated machine-learning techniques to identifying zone boundaries could not match the 90.15% recall achieved by this simple heuristic. The boundaries missed by the simple heuristic are all ADJOINING BOUNDARIES, where two zones are not separated by any buffer lines. An example of a boundary that is not detected by our heuristic is the zone boundary between the *Author* and *Signoff* zones in Figure 4.6 formed by lines 10 and 11.

Obviously, our simple boundary heuristic detects ACTUAL BOUNDARIES as well as SPURIOUS BOUNDARIES that do not actually separate different email zones. Unsurprisingly, the number of spurious boundaries is large. The precision of our simple heuristic across our annotated set of email messages is 22.5%, meaning that less than 1 in 4 hypothesised zone boundaries is an actual boundary. The underlying email zones average more than 12 lines in length, including just over 8 lines of non-blank content. Due to the number of spurious boundaries, fragments contain less than half this amount—approximately 3 lines of non-blank content on average. One of the most common types of spurious boundaries detected are the blank lines that frequently separate paragraphs within a single zone.

For three-zone classification, the set of predicted boundaries remains the same, but there are less actual boundaries to find, so recall increases to 96.3%. However, because many boundaries from the nine-zone classification are not boundaries for the three-zone classification, precision decreases to 14.7%.

Classifying Zone Fragments Having segmented the email message into candidate zone fragments, we classify these fragments using the SMO implementation provided by Weka (Witten and Frank, 2005) with the features described in Section 4.4.5.

Although our boundary detection heuristic has better than 90% recall, the small number of actual boundaries that are not detected result in some zone fragments containing lines from more than one underlying email zone. In these cases, we consider the mode of all annotation values for lines in the fragment (i.e., the most frequent zone annotation) to be the gold-standard zone type for the fragment. This, of course, may mean that we somewhat unfairly penalise the accuracy of our automated classification when Zebra detects a zone that is indeed present in the fragment, but is not the most frequent zone.

4.4.4.2 Line-based Zone Segmentation and Classification

Our line-based classification approach simply extracts all non-blank lines from an email message and classifies lines one-by-one, using the same features as for fragment-based classification. This approach is the same as the signature and reply line classification approach used by Carvalho and Cohen (2004).

4.4.5 Features for Zone Classification

We use a variety of graphic, orthographic and lexical features for classification in Zebra. The same features are applied in both the line-based and the fragment-based zone classification (to either individual lines or zone fragments). In the description of our features, we refer to both single lines and zone fragments (collections of contiguous lines) as TEXT FRAGMENTS.

Graphic Features Our graphic features capture information about the presentation and layout of text in an email message, independent of the actual words used. This information is a crucial source of information for identifying zones. Such information includes how the text is organised and ordered, as well as the 'shape' of the text. The specific features we employ are:

- The number of words in the text fragment [Integer];
- The number of Unicode code points (i.e., characters) in the text fragment [Integer];
- The start position of the text fragment (equal to one for the first line in the message, two for the second line and increasing monotonically through the message [Integer];

- The start position of the text fragment (as above) normalised for message length [Integer];
- The end position of the text fragment (calculated as above) [Integer];
- The end position of the text fragment (as above) normalised for message length [Integer];
- The average line length (in characters) within the text fragment (equal to the line length for line-based text fragments) [Integer];
- The length of the text fragment (in characters) relative to the previous fragment [Integer];
- The length of the text fragment (in characters) relative to the following fragment [Integer];
- The number of blank lines preceding the text fragment [Integer]; and
- The number of blank lines following the text fragment [Integer].

Orthographic Features Our orthographic features capture information about the use of distinctive characters or character sequences including punctuation, capital letters and numbers. Like our graphic features, orthographic features tend to be independent of the words used in an email message. The specific orthographic features we employ include:

- Whether all lines start with the same character (e.g., '>') [Boolean];
- Whether a prior text fragment in the message contains a quoted header [Boolean];
- Whether a prior text fragment in the message contains repeated punctuation characters [Boolean];
- Whether the text fragment contains a URL [Boolean];
- Whether the text fragment contains an email address [Boolean];
- Whether the text fragment contains a sequence of four or more digits [Boolean];
- The number of capitalised words in the text fragment [Integer];
- The percentage of capitalised words in the text fragment [Integer];
- The number of non-alpha-numeric characters in the text fragment [Integer];
- The percentage of non-alpha-numeric characters in the text fragment [Integer];
- The number of numeric characters in the text fragment [Integer];
- The percentage of numeric characters in the text fragment [Integer];
- Whether the message subject line contains a reply syntax marker such as *Re:* [Boolean]; and
- Whether the message subject line contains a forward syntax marker such as *Fw:* [Boolean].

Lexical Features Finally, our lexical features capture information about the words used in the email text. We use unigrams to capture information about the vocabulary and word bigram features to capture short range word order information. More specifically, the lexical features we apply to each text fragment include:

	2 Zones		3 Zones		9 Zones	
	Zebra	Baseline	Zebra	Baseline	Zebra	Baseline
Lines	93.60%	61.14%	91.53%	58.55%	87.01%	30.94%
Fragments	92.09%	62.18%	91.37%	59.44%	86.45%	30.36%

Table 4.6: Zone classification accuracy compared against a majority baseline.

	2 Zones		3 Zones		9 Zones	
	Zebra	Baseline	Zebra	Baseline	Zebra	Baseline
Lines	90.62%	61.14%	86.56%	58.55%	81.05%	30.94%
Fragments	91.14%	62.18%	89.44%	59.44%	82.55%	30.36%

Table 4.7: Zone classification accuracy, without word *n*-gram features, compared against a majority baseline.

- Each word unigram, calculated with a minimum frequency threshold cutoff of three, represented as a separate binary feature [Boolean];
- Each word bigram, calculated with a minimum frequency threshold cutoff of three, represented as a separate binary feature [Boolean];
- Whether the text fragment contains the sender's name [Boolean];
- Whether a prior text fragment in the message contains the sender's name [Boolean];
- Whether the text fragment contains the sender's initials [Boolean]; and
- Whether the text fragment contains a recipient's name [Boolean].

Features that look for instances of sender or recipient names are less likely to be specific to a particular business or email domain. These features use regular expressions to find name occurrences, based on semi-structured information in the email message headers. First, we extract and normalise the names from the email headers to identify the relevant person's given name and surname. Our features then capture whether one or both of the given name or surname are present in the current text fragment. Features which detect user initials make use of the same name normalisation code to retrieve a canonical form of the user's name, from which their initials are derived.

4.4.6 Intrinsic Evaluation of Email Zone Segmentation and Classification

Table 4.6 shows Zebra's accuracy in classifying email zones. The results are calculated using stratified, 10-fold cross validation. Accuracy is shown for three tasks—nine-,

	Total	Au	Si	Di	Ad	Gr	So	Re	Fw	At
Author (Au)	2415	2197	56	9	4	14	31	43	53	8
Signature (Si)	383	93	203	4	0	0	20	28	31	4
Disclaim (Di)	97	30	4	52	0	0	0	2	9	0
Advert (Ad)	83	47	1	1	20	0	0	7	7	0
Greet (Gr)	85	8	0	0	0	74	2	0	1	0
Signoff (So)	195	30	5	0	0	0	147	11	2	0
Reply (Re)	2451	49	10	3	2	1	10	2222	154	0
Fwd (Fw)	2187	72	13	7	8	1	3	125	1958	0
Attach (At)	26	4	0	0	0	0	0	1	1	20

Table 4.8: The confusion matrix for nine-zone, line-based classification.

three- and two-zone classification—using both line and zone-fragment classification. Performance is compared against a majority class baseline in each case.

Zebra's performance compares favourably with previously published results. While it is difficult to directly compare, since not all systems are freely available and they are not trained or tested over the same data, our three-zone classification (identifying sender, boilerplate and quoted reply content) is very similar to the three-zone task for which (Estival et al., 2007) report 88.16% accuracy for their system and 64.22% accuracy using Carvalho and Cohen's Jangada system. Zebra outperforms both, achieving 91.53% accuracy using a line-based approach. In the two-zone task, where we attempt to identify sender-authored lines, Zebra achieves 93.60% accuracy and an F-score of 0.918, exceeding the 0.907 F-score reported for Estival et al.'s system tuned for exactly this task.

Interestingly, the line-based approach provides slightly better performance than the fragment-based approach for each of the two-zone, three-zone and nine-zone classification tasks. As noted earlier, our original hypothesis was that zone fragments would contain more information about the sequence and text shape of the original message, and that this would lead to better performance for fragment-based classification.

When we restrict our feature set to those that look only at the text of the line or zone fragment, the fragment-based approach does perform better than the line-based one. Using only word unigram features, for example, our fragment classifier achieves 78.7% accuracy. Using the same features, the line-based classifier achieves only 57.5% accuracy. When we add further features that capture sequence and shape information from outside the text fragment being classified (e.g., the length of a text segment compared to the text segment before and after, and whether a segment occurs after another segment containing repeated punctuation or the sender's name), the line-based approach achieves a greater increase in accuracy than the fragment-based approach. This presumably is because individual lines intrinsically have less information about the message context, and so benefit more from the information added by the new features.

We also experimented with removing all word unigram and bigram features to explore the classifier's portability across different domains. This removed all vocabulary and word order information from our feature set. In doing so, our feature set was reduced to less than thirty features, consisting of mostly graphic and orthographic information. The few remaining lexical features captured only the presence of sender and recipient names, which are independent of any particular email domain. As expected, performance did drop, but not dramatically. Table 4.7 shows that average performance without n-grams (across two-, three- and nine-zone tasks) for line-based classification drops by 4.67%. In contrast, fragment-based classification accuracy drops by less than half this amount—an average of 2.26%. This suggests that, as we originally hypothesised, there are additional non-lexical cues in zone fragments that give information about the zone type. This makes the zone fragment approach potentially more portable for use across email data from different enterprise domains.

Of course, classification accuracy gives only a limited picture of Zebra's performance. Table 4.9 shows precision and recall results for each zone in the nine-zone line-based classification task. Performance clearly varies significantly across the different zones. For *Author*, *Greeting*, *Reply* and *Forward* zones, performance is good, with F-score greater than 0.8. This is encouraging, given that many email tools, such as action-item detection and email summarisation would benefit from an ability to separate author content from reply content and forwarded content. The *Advertising*, *Signature* and *Disclaimer* zones show the poorest performance, particularly in terms of Recall. The *Advertising* and *Disclaimer* zones are almost certainly hindered by a lack of training data; they are two of the smallest zones in terms of number of lines of training data. The relatively poor *Signature* class performance is more interesting. Given the potential confusion between *Signoff* content and *Signature*s that function as *Signoff*s, one might expect confusion between *Signoff* and *Signature* zones, but Table 4.8 shows this is not the case. Instead, there is significant confusion between *Signature* and *Author* content, with almost 25% of *Signature* lines misclassified as *Author* lines. When word n-grams are removed from the feature set, the number of these misclassifications increases to almost 50%. These results reinforce our observation that the task of email signature extraction is much more difficult that it was in the days of Usenet messages.

4.5 Improving Obligation Act Classification with Email Zoning

Requests and commitments in email do not occur uniformly across the zones that make up the email message. There are specific zones of a message in which these obligation acts are likely to occur. Indeed, requests and commitments that occur in some message zones may almost always not be actionable, regardless of the utterance content. For example, calls-to-action and rhetorical questions occur frequently in advertising content, such as that shown in the highlighted section of Figure 4.7. Although they

Zone	Precision	Recall	F-score
Author	0.868	0.910	0.889
Signature	0.695	0.530	0.601
Disclaimer	0.684	0.536	0.601
Advertising	0.588	0.241	0.342
Greeting	0.822	0.871	0.846
Signoff	0.690	0.754	0.721
Reply	0.911	0.907	0.909
Forward	0.884	0.895	0.889
Attachment	0.625	0.769	0.690

Table 4.9: Precision and recall for nine-zone, line-based zone classification.

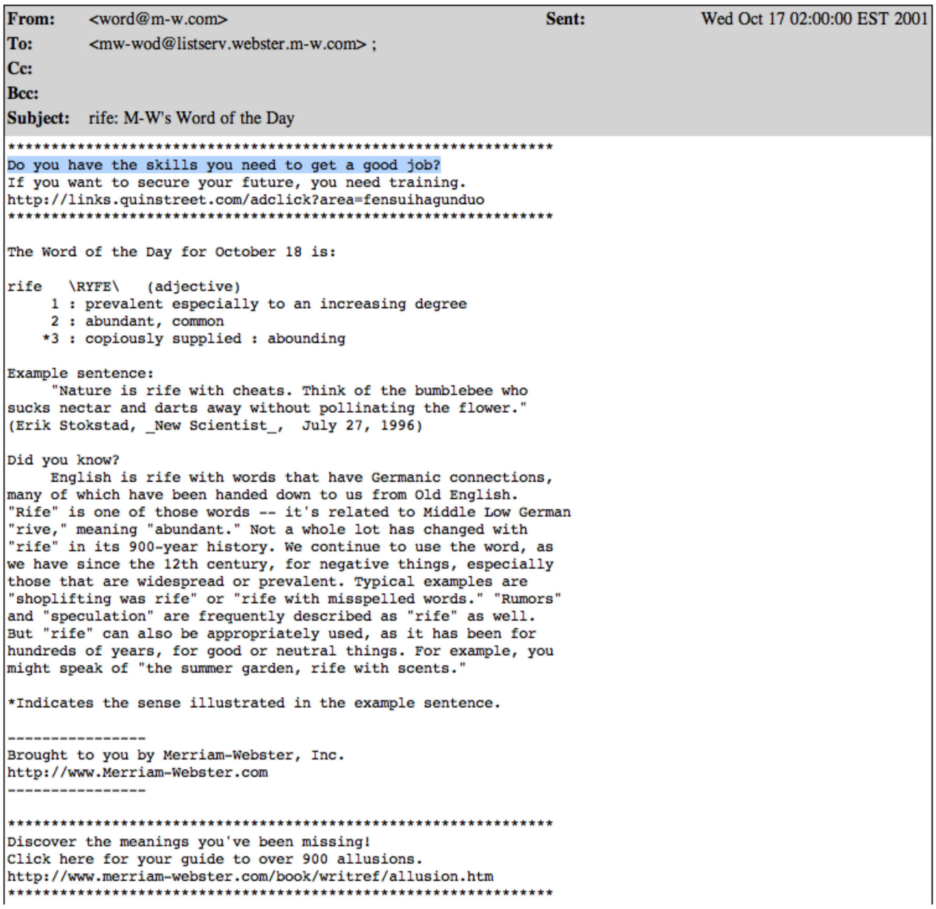
resemble the form of requests, such utterances would rarely be considered to convey obligation for a recipient to act or respond. We employ the Zebra system described in Section 4.4 to test our hypothesis that we can improve request and commitment classification performance by first classifying the body text of each email.

Despite the likelihood of some noise being introduced through misclassification of email zones, our hypothesis was that even imperfect information about the functional parts of a message should improve the performance of our request classifier.

Based on this hypothesis, we integrated Zebra to identify the different functional parts of email messages. To make our processing pipeline simpler, we configured Zebra for line-based zone classification, and use it to extract only lines classified as author, greeting and signoff text. We remove the content of all other zones before we evaluate features for obligation act classification.

4.5.1 Request Classification with Email Zoning

Classifying the zones in email messages and then applying our request classifier to only relevant message parts significantly increases the performance of the request classifier. As detailed in Section 4.3.1, without zoning, our request classifier achieves accuracy of 72.28% and a weighted F-score (weighted between the F-score for requests and non-requests based on the relative frequency of each class) of 0.723. Adding the zone classifier, we increase the accuracy to 83.76% and the weighted F-score to 0.838. This corresponds to a relative increase in both accuracy and weighted F-score of 15.9%, which in turn corresponds to an error reduction of more than 41%. Table 4.10 shows a comparison of the results of the non-zoning and zoning request classifiers, generated using stratified 10-fold cross validation. In a two-tailed paired t-test, run over ten iterations of stratified 10-fold cross validation, the increase in accuracy, precision, recall and F-score were all significant at p=0.01.



From: <word@m-w.com> **Sent:** Wed Oct 17 02:00:00 EST 2001
To: <mw-wod@listserv.webster.m-w.com> ;
Cc:
Bcc:
Subject: rife: M-W's Word of the Day

```
***************************************************************
Do you have the skills you need to get a good job?
If you want to secure your future, you need training.
http://links.quinstreet.com/adclick?area=fensuihagunduo
***************************************************************

The Word of the Day for October 18 is:

rife   \RYFE\   (adjective)
     1 : prevalent especially to an increasing degree
     2 : abundant, common
    *3 : copiously supplied : abounding

Example sentence:
     "Nature is rife with cheats. Think of the bumblebee who
sucks nectar and darts away without pollinating the flower."
(Erik Stokstad, _New Scientist_,  July 27, 1996)

Did you know?
     English is rife with words that have Germanic connections,
many of which have been handed down to us from Old English.
"Rife" is one of those words -- it's related to Middle Low German
"rive," meaning "abundant." Not a whole lot has changed with
"rife" in its 900-year history. We continue to use the word, as
we have since the 12th century, for negative things, especially
those that are widespread or prevalent. Typical examples are
"shoplifting was rife" or "rife with misspelled words." "Rumors"
and "speculation" are frequently described as "rife" as well.
But "rife" can also be appropriately used, as it has been for
hundreds of years, for good or neutral things. For example, you
might speak of "the summer garden, rife with scents."

*Indicates the sense illustrated in the example sentence.

----------------
Brought to you by Merriam-Webster, Inc.
http://www.Merriam-Webster.com
----------------

***************************************************************
Discover the meanings you've been missing!
Click here for your guide to over 900 allusions.
http://www.merriam-webster.com/book/writref/allusion.htm
***************************************************************
```

FIGURE 4.7: An email message containing highlighted request-like utterances in an advertising zone.

4.5.1.1 Lexical Feature Contribution

As expected, lexical information remains crucial to request classification. When we previously experimented with removing all lexical (n-gram) features, the non-zoning request classifier accuracy dropped to 57.62%. Our classifier with zoning also suffers a significant drop in accuracy this time to 61.78%. In relative terms, the drop is similar, but clearly zoning the messages first makes some of our non-lexical features (such as the presence of sender or recipient names) more salient than before.

A bigger contrast is evident, when we apply *only* n-gram features. Our non-zoning request classifier achieved accuracy of 71.49% while our request classifier with zoning now achieves 83.36% accuracy. Clearly, lexical information is critical for accurate request classification, regardless of whether email messages are zoned, but zoning the message text first appears to increase the reliability of lexical cues. This is as expected, as ignoring irrelevant text reduces the likelihood of erroneously identifying a request in message zones such as email signatures and advertising. Advertising content in

	No Zoning		With Zoning	
	Request	Non-Request	Request	Non-Request
Accuracy	72.28%		83.76%*	
Precision	0.729	0.716	0.849*	0.825*
Recall	0.745	0.698	0.837*	0.839*
F-score	0.737	0.707	0.843*	0.832*

Table 4.10: Request classifier results with and without email zoning (* indicates a statistically significant difference at p=0.01).

particular frequently contains directive utterances, in an attempt to persuade readers to act or purchase, as we saw earlier in Figure 4.7.

We also see a marked contrast in the top-10 unigrams and bigrams, as ranked by Information Gain, for the request classifier when coupled with our zone classifier. When compared against the top-10 unigrams and bigrams for our initial request classifier, presented in Table 4.3, the results for our zoning request classifier, shown in Table 4.11, appear to correspond much better with linguistic intuitions about the language of requests. Using only these top-10 n-grams (plus our non-n-gram features), we achieve 80% accuracy. This suggests that, even with our relatively small amount of training data, the zone classifier helps the request classifier to extract fairly general n-gram features.

Although lexical features are very important, the top three features ranked by Information Gain are non-lexical: message length in words, the number of non-alphanumeric characters in the message and the number of capitalised words in the message. All these tend to trend towards longer messages being more likely to contain a request. This illustrates that our combination of lexical and non-lexical cues are both contributing to the success of our request classifier.

4.5.1.2 Learning Curves

Figure 4.8 shows a plot of accuracy, precision and recall *versus* the number of training instances used to build the request classifier. These results are calculated over zoned email bodies, using the average across ten iterations of stratified 10-fold cross validation for each different sized set of training instances, implemented via the `FilteredClassifier` with the `Resample` filter in Weka. Given our pool of 505 agreed message annotations, we plot the recall and precision for training instance sets of size 50 to 505 messages.

Unsurprisingly, there is a clear trend of increasing performance as the training set size grows. It seems reasonable to assume that more data should continue to facilitate better request classifier performance. This provides scope for further improvement in

Word Unigrams	Word Bigrams	
	Word 1	Word 2
please	?	END-SENTENCE
?	PRONOUN-OBJECT	know
PRONOUN-ACCUSATIVE	let	PRONOUN-ACCUSATIVE
if	START-SENTENCE	please
PRONOUN-ACCUSATIVE	if	PRONOUN-NOMINATIVE
let	START-SENTENCE	thanks
to	please	let
know	PRONOUN-NOMINATIVE	have
thanks	thanks	COMMA
do	start	date

Table 4.11: The top 10 useful n-grams for our request classifier with zoning, ranked by Information Gain.

performance through the collection of more annotated data, suggesting that our approach will scale beyond the limited scope of annotated datasets that we have curated.

4.5.1.3 Error Analysis

To explore the errors made by our request classifier, we examined the output of our request classifier coupled with Zebra, using its full feature set.

Approximately 20% of errors relate to requests that are implicit, and thus more difficult to detect from surface features. Another 10% of errors are due to attempts to classify requests in inappropriate genres of email messages. In particular, both marketing messages and spam frequently include request-like, directive utterances which our annotators all agreed would not be useful to mark as requests for an email user. Not unreasonably, our classifier is sometimes confused by the content of these messages, mistakenly marking requests where our annotators did not. An example of a message that was identified by our classifier, but not marked by our human annotators, is shown in Figure 4.9. One way to resolve these classification errors would be to filter out messages of particular genres before we apply the request classifier.

Another 5% of errors are due to request content occurring in zones that we ignore. The most common case is content in a forwarded zone. Sometimes email senders forward a message as a form of task delegation; because we ignore forwarded content, our request classifier misses such requests. We discussed this issue more completely in Section 3.3.4, and presented an example back in Figure 3.27.

We experimented with including content from forwarded zones (in addition to the author, greeting and signoff zones), but found that this reduced the performance of our request classifier, presumably due to the additional noise from irrelevant content in other forwarded material. Forwarded messages are thus somewhat difficult to classify.

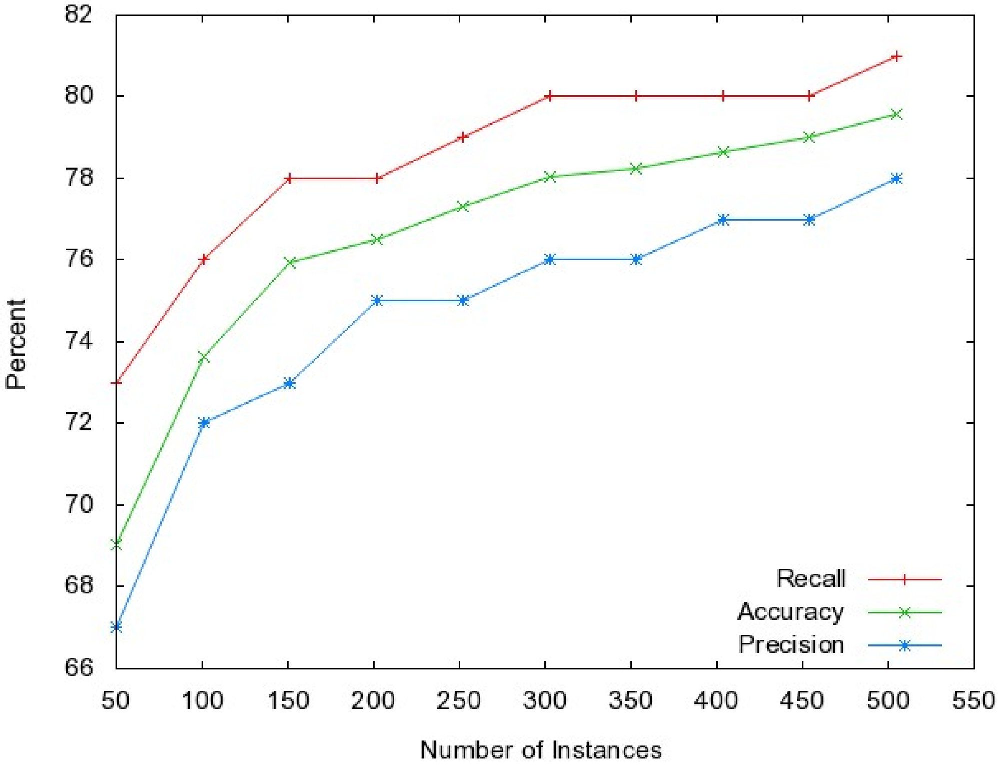
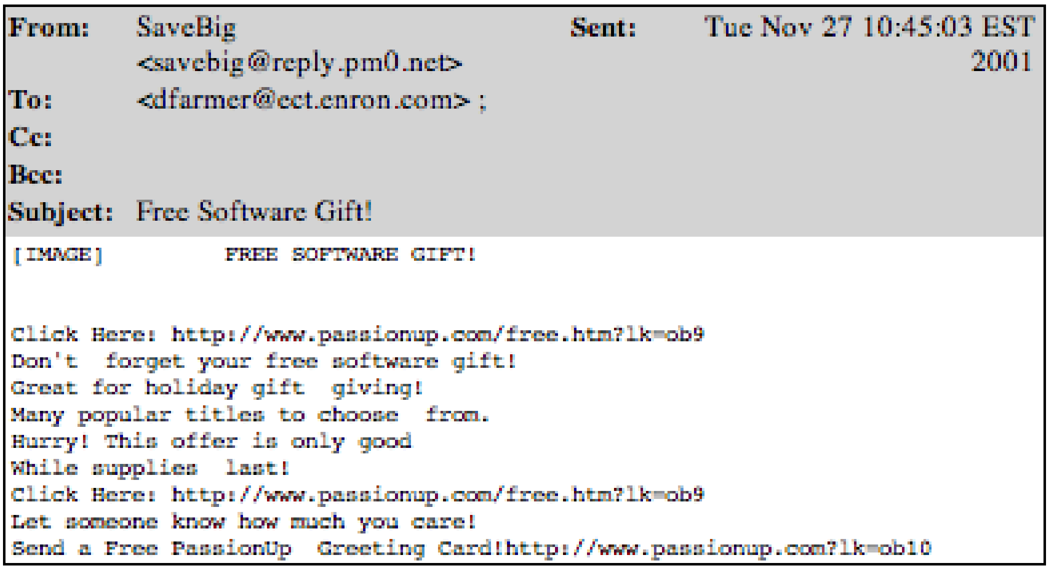


Figure 4.8: The learning curve for our request classifier with zoning, showing recall, accuracy and precision *versus* the number of training instances.

One possible approach would be to build sender-specific profiles that might allow us to deal with forwarded content (and potentially content from other zones) differently for different users, essentially learning to adapt to the different styles of different email users.

A further 5% of errors involve errors in the zone classifier, which leads to incorrect zone labels being applied to zone content that we would wish to include for our request classifier. Examples include author content being mistakenly identified as signature content. In such cases, we incorrectly remove relevant content from the body text that is passed to our request classifier. Improvements to the zone classifier would resolve these issues.

As part of our annotation task, we also asked annotators to mark the presence of pleasantries, which we now refer to as phatic requests and phatic commitments. As we discussed in Section 3.3.1, phatic obligation acts occur where an utterance that could function as a request or a commitment in some other context, but that do not convey obligation in the context of use under consideration. Phatic acts in email are frequently formulaic, and do not place any significant obligation on the recipient to act or respond. Variations on the phrase *Let me know if you have any questions* are particularly common. The context of the entire email message needs to



From: SaveBig <savebig@reply.pm0.net> **Sent:** Tue Nov 27 10:45:03 EST 2001
To: <dfarmer@ect.enron.com> ;
Cc:
Bcc:
Subject: Free Software Gift!

```
[IMAGE]          FREE SOFTWARE GIFT!


Click Here: http://www.passionup.com/free.htm?lk=ob9
Don't  forget your free software gift!
Great for holiday gift  giving!
Many popular titles to choose  from.
Hurry! This offer is only good
While supplies  last!
Click Here: http://www.passionup.com/free.htm?lk=ob9
Let someone know how much you care!
Send a Free PassionUp  Greeting Card!http://www.passionup.com?lk=ob10
```

Figure 4.9: An email message containing request-like phrases that were not marked by human annotators due to the marketing/spam nature of the email message.

be considered to distinguish between when such an utterance functions as a request or a commitment, and when it should be marked as a phatic act. Of the errors made by our request classifier, approximately 5% involve marking messages containing only phatic requests as containing a request.

A little less than 5% of errors involve problems interpreting requests associated with attached files. In some cases, it is clear than including an attachment conveys a request to read or act on that material, however, in other cases it is less clear, leading to ambiguity. We discussed this issue in more detail in Section 3.3.9.

The remaining errors are somewhat diverse. The balance of almost 50% of total errors involve a wide range of issues, from misspellings of key words such as PLEASE to a lack of punctuation cues such as question marks.

4.5.2 Commitment Classification with Email Zoning

As discussed in Section 4.3.2, our initial commitment classifier, without zoning, achieves overall accuracy of 82.77% and a weighted F-score (weighted between the F-score for commitments and non-commitments based on the relative frequency of each class) of 0.799. The headline result from adding our Zebra zone classifier to filter irrelevant sections of text from the messages before feature generation for commitment classification is that we increase the commitment detection accuracy at the message-level to 86.62% and the weighted F-score to 0.858.

This corresponds to a relative increase in accuracy of 4.6% and weighted F-score of 7.4%, which in turn corresponds to an error reduction of 22.3%. Table 4.12 shows a comparison of the results between our zoned and non-zoned commitment classifiers.

	No Zoning		With Zoning	
	Commit	Non-Commit	Commit	Non-Commit
Accuracy	82.77%		86.62%*	
Precision	0.809	0.830	0.807	0.878*
Recall	0.357	0.978	0.571*	0.958
F-score	0.495	0.896	0.669*	0.916*

Table 4.12: Commitment classifier results with and without email zoning (* indicates a statistically significant difference at p=0.01).

Word Unigrams	Word Bigrams	
	Word 1	Word 2
PRONOUN-NOMINATIVE	PRONOUN-NOMINATIVE	will
will	would	like
be	PRONOUN-NOMINATIVE	would
if	will	be
week	like	to
discuss	if	PRONOUN-NOMINATIVE
tomorrow	next	week
let	PRONOUN-ACCUSATIVE	to
I'll	to	discuss
next	let	PRONOUN-ACCUSATIVE

Table 4.13: The top 10 useful n-grams for our commitment classifier without zoning, ranked by Information Gain.

Again, results are all generated using stratified 10-fold cross validation. In two-tailed paired t-tests over ten iterations of stratified 10-fold cross validation, the increase in accuracy, recall and F-score were all significant at $p = 0.01$. The differences in weighted F-score (0.799 *vs.* 0.858) were also significantly different at $p = 0.01$

Table 4.13 shows the top-10 n-grams for our zoning commitment classifier. Unlike for our request classifier, there is not a strong difference in the top features when Zebra is applied before commitment classification. This suggests that there is less confusion from commitments in irrelevantly zoned text than there is for requests. Intuitively, this makes sense, as content that is zoned to be ignored, such as signatures and legal disclaimers, more frequently contain requests and than commitments. The increased performance from adding zoning, however, clearly must influence the efficacy of the remaining n-gram and non-n-gram features, given the increase in accuracy.

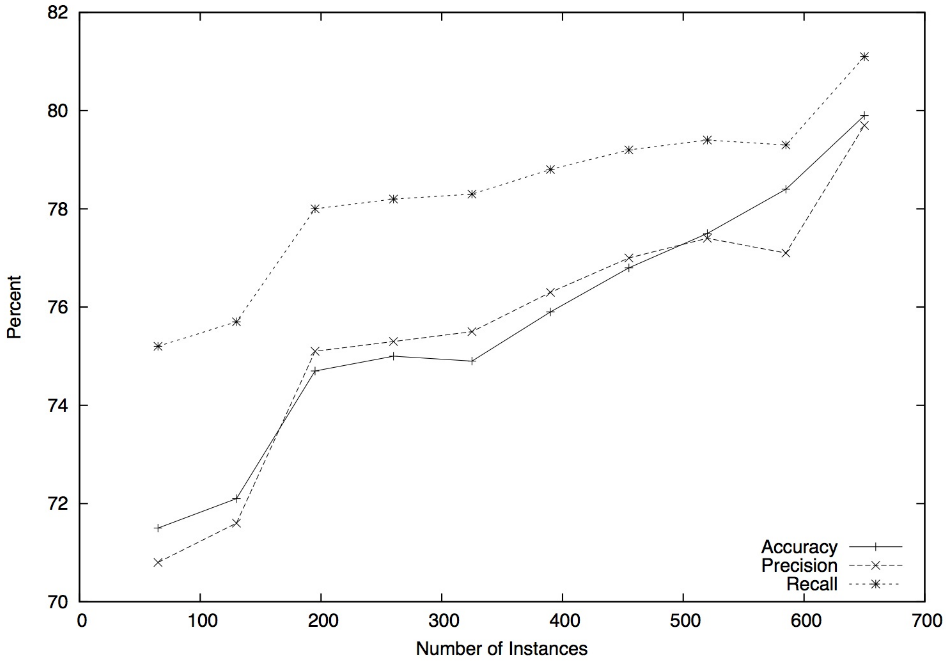


Figure 4.10: The learning curve for our commitment classifier with zoning, showing recall, accuracy and precision *versus* the number of training instances.

4.5.2.1 Learning Curves

Figure 4.10 shows a plot of accuracy, precision and recall *versus* the number of training instances used to build our zoning commitment classifier. These results are calculated over zoned email bodies, using the average across ten iterations of stratified 10-fold cross validation for each different sized set of training instances, implemented via the `FilteredClassifier` with the `Resample` filter in Weka. Given our pool of 650 unanimously agreed message annotations, we plot the recall and precision for training instance sets of size 65 to 650 messages, in 10% increments.

There is a relatively stable trend of increasing performance as the training set size grows, but this is not monotonic, as in the case of our request classifier. It is likely that further annotated data would lead to continued improvement in classification accuracy, however, it is difficult to be confident of this given the fluctuations in performance. The variance suggests that our training set may contain artefacts in the training/test splits that are disproportionately influencing the results for our different training set sizes.

4.5.2.2 Error Analysis

To explore the errors made by our commitment classifier, we examined the output of our zoning commitment classifier using our full feature set, including all word n-grams.

Approximately 30% of errors relate to commitments that are implicit, and thus more difficult to detect from surface features. Commitments are more challenging to recognise than requests in this regard. Take, for example, the following utterance:

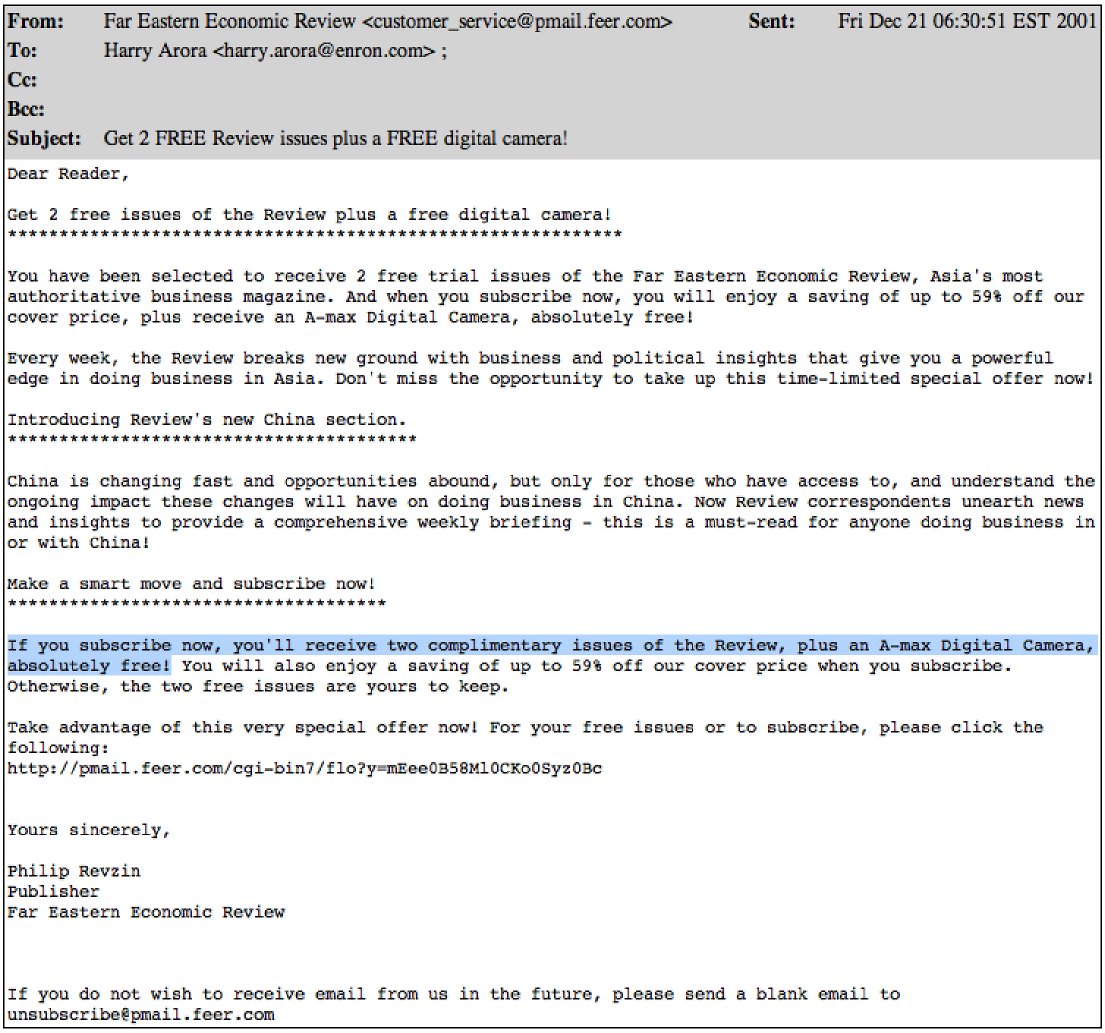
(4.3) *The most intense heat will continue in the southwestern states, but the hot weather will move into Texas and portions of the central Plains by early next week, further boosting demand.*

While this utterance has the syntactic form of a commitment, its meaning is actually closer to a prediction; the weather is something that the author of this utterance clearly does not have control over and cannot be reasonably be seen to be taking on an obligation to ensure hot weather moves into Texas. The distinction between predictions and commitments is frequently difficult to discern from the syntactic forms alone, meaning that the lexical cues that can be used for classification are not as consistent as those for requests.

As for requests, approximately 5–10% of errors are due to attempts to classify commitments in inappropriate genres of email messages. In particular, both marketing messages and spam frequently include commitment-like utterances which our annotators all agreed would not be useful to mark as commitments for an email user to follow up on. An example is shown in Figure 4.11. Not unreasonably, our classifier is sometimes confused by the content of these messages, identifying commitments where our annotators did not. As for requests, one way to resolve classification errors of this type would be to filter out such messages before we apply the commitment classifier.

Less than 5% of commitment classification errors involve errors in the zone classifier, which leads to incorrect zone labels being applied to zone content that we would wish to include for our commitment classifier. Examples include author content being mistakenly identified as signature content. In such cases, we incorrectly remove relevant content from the body text that is passed to our commitment classifier. As for requests, improvements to the zone classifier would be the most obvious way to resolve this class of errors; however, the low frequency of these errors adds to the evidence from our request classification experiments that further improvements to our zone classifier will have diminishing returns for our obligation act classification.

The remaining errors are quite diverse, involving a wide range of issues, including ambiguity and implicitness. There were also a small number of cases where commitments that had been identified by our human annotators were in zones that were correctly ignored, notably in forwarded content. This is the same issue that we have seen arise for request classification and is difficult to resolve, as forwarded content varies as to whether the sender intends to place obligations for action from forwarded requests and commitments on the new recipient.



From: Far Eastern Economic Review <customer_service@pmail.feer.com> **Sent:** Fri Dec 21 06:30:51 EST 2001
To: Harry Arora <harry.arora@enron.com> ;
Cc:
Bcc:
Subject: Get 2 FREE Review issues plus a FREE digital camera!

Dear Reader,

Get 2 free issues of the Review plus a free digital camera!
****

You have been selected to receive 2 free trial issues of the Far Eastern Economic Review, Asia's most authoritative business magazine. And when you subscribe now, you will enjoy a saving of up to 59% off our cover price, plus receive an A-max Digital Camera, absolutely free!

Every week, the Review breaks new ground with business and political insights that give you a powerful edge in doing business in Asia. Don't miss the opportunity to take up this time-limited special offer now!

Introducing Review's new China section.
****

China is changing fast and opportunities abound, but only for those who have access to, and understand the ongoing impact these changes will have on doing business in China. Now Review correspondents unearth news and insights to provide a comprehensive weekly briefing - this is a must-read for anyone doing business in or with China!

Make a smart move and subscribe now!
****

If you subscribe now, you'll receive two complimentary issues of the Review, plus an A-max Digital Camera, absolutely free! You will also enjoy a saving of up to 59% off our cover price when you subscribe. Otherwise, the two free issues are yours to keep.

Take advantage of this very special offer now! For your free issues or to subscribe, please click the following:
http://pmail.feer.com/cgi-bin7/flo?y=mEee0B58Ml0CKo0Syz0Bc

Yours sincerely,

Philip Revzin
Publisher
Far Eastern Economic Review

If you do not wish to receive email from us in the future, please send a blank email to unsubscribe@pmail.feer.com

Figure 4.11: An email message containing commitment-like phrases that were identified by our commitment classifier, but not marked by human annotators due to the marketing/spam nature of the email message.

4.6 Summary

This chapter has described how we have built two message-level text classifiers that identify whether an email message conveys an obligation act for the sender or recipient. Request and commitment classification, like any form of automated speech act recognition, is a difficult task due to the indirect relationship between lexical form and speech act function.

Unlike previous work that has attempted to automate the classification of requests or commitments in email, we introduce the novel step of zoning the email messages without manual intervention prior to request and commitment classification. Our hypothesis was that doing so would improve the performance of request and commitment classifiers. This hypothesis has been confirmed. The automatic classifiers that we describe in this chapter, when coupled with our automated email zoning system, correctly identify requests and commitments at the message level in 83.76% and 86.62% of email messages respectively. Specifically, we found that employing Zebra, our automated email zone classification system, to focus our obligation act classifiers on relevant parts of each message improves accuracy by between 4.6% and 15.9%, compared with the performance of the same classifiers operating without the assistance of an email zone classifier. Although some zone classification errors are made, our error analysis reveals that only approximately 5% of errors are due to zone misclassification of message parts for both request and commitment classification. This suggests that, although there is clearly headroom to improve zone classifier performance, it is likely that focusing instead on improving the request and commitment classifiers using the existing zone classifier performance would lead to greater performance gains.

For both request and commitment classification, lexical features provide the strongest cues. These cues are made more reliable through the coupling of Zebra, our zone classifier, to filter message text before classification. We leverage this knowledge in the design of our finer-grained request and commitment classifiers in Chapter 5.

"(I) automatically keep mails when I have sent them or received them so I know that I have got a record. For example, 'you sent me an email saying that you'll do this so ..."

Quote from (Brown, Middleton, and Lightfoot, 2001, p. 135)

5 Fine-grained Classification of Requests and Commitments

The message-level classifiers we developed in Chapter 4 can be applied in email software to focus a user's attention on email messages which contain actionable content, thus helping to triage messages in their inbox. To further support email task management, we also envisage future services that can automatically summarise the actionable content in each email message or automatically extract requests and commitments into a user's structured task list. Such services require requests and commitments to be identified at a finer granularity than the message level.

In this chapter we develop and evaluate techniques for identifying requests and commitments at the paragraph level and sentence level. We choose paragraphs as a unit that is likely to include sufficient context around each request or commitment utterance that they can be extracted and presented, independently from the original message, to summarise or collate obligations across a collection of messages. Sentence-level identification is chosen as the finest granularity for which we could achieve acceptable human annotator agreement in our data. Clause-based or text fragment classification is left to future work.

Building on our message-level classification work, we approach this finer-grained identification task using supervised machine learning methods within a text classification framework. In Section 5.1, we describe how we derive our separate, labelled corpora of paragraphs and sentences from our corpus of variable-length text spans from 1000 business email messages that have been annotated by two independent annotators. We have made these corpora of labelled paragraphs and sentences available publicly for use by other researchers.

In Section 5.2, we present our approach to fine-grained classification, built upon Support Vector Machine (SVM) classifiers with a range of lexical and syntactic features. We also describe how we deal with the increased data skew observed in our annotated data at these finer granularities, specifically through the application of under-sampling, over-sampling and cost-sensitive classification, and discuss how we evaluate classification performance.

Section 5.3 then describes and evaluates our fine-grained request classifiers. We discuss the features employed and present a detailed evaluation and analysis of classification performance over our derived corpora of request-annotated paragraphs and sentences. We also perform experiments with the techniques described in Section 5.2 to combat imbalance in our labelled datasets, and discover that their applicability is limited for the relatively low levels of data skew seen in our request corpora. Approximately 21–30% of paragraphs and sentences contain requests.

We then describe and evaluate an equivalent set of commitment classifiers in Section 5.4, again operating at the paragraph and sentence levels. Data skew is a much more significant problem for these classifiers, with only 4–10% of paragraphs and sentences containing a commitment. We present a detailed evaluation of the efficacy of under-sampling and cost-sensitive classification in improving the results of our baseline SVM classifiers for paragraph and sentence classification.

In general, our attempts to combat data skew through the use of under-sampling and cost-sensitive classification do not lead to a consistent improvement in performance. To explore alternate ways of improving performance for our commitment classifiers, we design and evaluate two ensemble approaches for combating fine-grained classification errors in Section 5.5. These coarse-to-fine techniques are built on the recognition that our message, paragraph and sentence classifiers all identify the same phenomena at differing levels of detail. We leverage this insight by combining our classifiers at different granularities into classifier ensembles. The ensemble technique, which augments the feature space of finer-granularity classifiers, is able to provide significant improvement over the standalone performance for both paragraph and sentence classification.

Overall, this chapter adapts our message-level classification approach to identify requests and commitments at finer levels of granularity. We present and evaluate a series of independent and ensemble classification techniques for identifying actionable content in email messages at the paragraph and sentence level. Our resulting classifiers provide the required functionality for the experimental email software integration that we describe in Chapter 6. There, we embed our collection of request and commitment classifiers at different granularities within Microsoft Outlook, both to provide useful tools for users working with tasks in email and to create a platform for further research and evaluation of tools and techniques for task management in email.

5.1 Creating Labelled Corpora of Paragraphs and Sentences

In order to employ supervised machine learning techniques, we require labelled request and commitment corpora to train and evaluate both our paragraph and sentence classifiers. Our review of related literature suggests that no such corpus has previously been made publicly available. We therefore created our own labelled corpora of paragraphs and sentences for both requests and commitments. This corpus has been made publicly available for other researchers to use.[1]

In Section 3.2.4, we described how we undertook our *Span 2* experiment which involved two independent annotators marking all request and commitment spans in 1000 email messages. In this section we describe how we transform the resulting corpus of identified and labelled text spans into separate corpora of annotated paragraphs and labelled sentences. This involves two main steps. Firstly, we segment the full text of each message body into paragraphs or sentences, as appropriate, and use gold-standard zone labels to identify the correct zone labels for each paragraph/sentence. We then discard paragraphs or sentences that do not occur in a relevant email zone. One implication of this is that we discard annotations that occur in an irrelevant zone. In practice, this occurs infrequently in our annotated corpus. Secondly, we generate a gold-standard request and commitment label for each remaining paragraph/sentence, based on the spans that have been annotated that overlap with the paragraph or sentence. There is significant complexity within these processes, so we expand on each of these steps below, including how we compute span overlaps, to explain exactly how we generate our labelled paragraph corpus.

5.1.1 Segmenting and Selecting Relevant Paragraphs

The first step in generating our labelled paragraph corpus is to segment the message body of the email into paragraphs. Mostly, paragraphs are separated by whitespace; indenting the first line of a paragraph, which is a common marker in written text, is extremely rare in email. One challenge when interpreting whitespace as a paragraph separator is that whitespace can work at different levels of a document's structural hierarchy, and is not marked in any way to indicate the level at which it operates; consequently, the function of whitespace in an email message is ambiguous. Figure 5.1 shows an example of whitespace operating at different levels within a message; some whitespace indicates a paragraph boundary while other whitespace operates within a paragraph as a separator between the list introduction and the first item in the list.

As part of our email zone classification work detailed in Section 4.4, we developed techniques for segmenting email body text into ZONE FRAGMENTS, which represent contiguous lines of email body text separated by lines containing only whitespace or punctuation. The ambiguity of whitespace means that it is possible for a paragraph to span multiple underlying zone fragments, as illustrated in Figure 5.1, or for a zone fragment to span multiple paragraphs, as shown in Figure 5.2. Despite this ambiguity, in many cases this very simple notion of a zone fragment corresponds well to a paragraph. Throughout this thesis we make the simplifying assumption that each zone fragment maps to exactly one paragraph for the purposes of paragraph classification. We rely on the zone boundary detection mechanisms described in Section 4.4.4 to find the boundaries between paragraphs.

Although we see examples of paragraphs spanning multiple zone fragments, each with the same zone label, none of our email corpora reveal any examples where a

[1] The corpus is available from `http://zebra.thoughtlets.org/spans/data.php`.

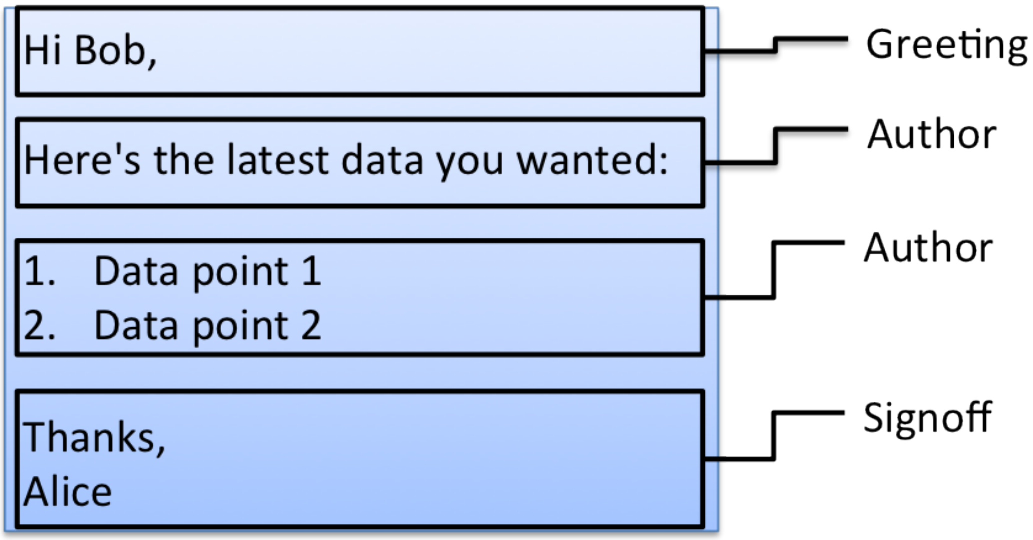


Figure 5.1: A synthesised example of where a single paragraph spans more than one zone fragment. In this case, the text covered by the two *Author* zones is arguably part of a single paragraph.

single paragraph spans across multiple zone fragments with different zone labels. Intuitively, this is an expected result; the topical cohesion of a typical paragraph makes the probability of spanning multiple functional zones of an email message quite low.

Where the reverse occurs, that is, where a single zone fragment spans multiple paragraphs, we treat the entire zone fragment as a single paragraph. An example of how this can occur is shown in Figure 5.2. In such cases, it is possible that the zone segment would have different zone annotations; our segmentation code is set up to create a separate instance of each identified paragraph for each gold-standard zone associated with a paragraph.

After segmenting the text of an email into paragraphs, we keep only the paragraphs from relevant zones of the email message. As in Chapter 4, relevant paragraphs are those marked as: Author Content; Signoff Content; or Attachment Content.[2]

Because we align our notion of zone fragments and paragraphs, we simply consider the gold-standard zone annotation associated with each paragraph to decide on its relevance. In the small number of cases where a single paragraph has multiple zone annotations, we keep exactly one instance of the paragraph if at least one of the gold-standard zone annotations for that paragraph is a relevant zone.

[2] A reminder from our work presented in Chapter 4 that Attachment content refers to text within the message body that refers to attached information; we do not have access to the attached content in the version of the Enron corpus with which we work.

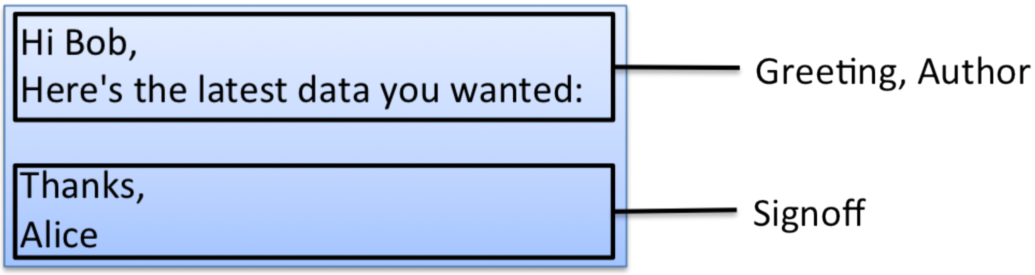


Figure 5.2: A synthesised example of where a single zone fragment spans more than one paragraph. In this case, the first zone fragment spans two different zones across the first two paragraphs of the message (*Hi Bob* and *Here's the latest data you wanted.*).

5.1.2 Generating a Gold-Standard Labelled Paragraph Corpus from Annotated Spans

After segmenting the paragraphs and discarding those from irrelevant zones, we generate gold-standard, binary request and commitment labels for each paragraph that indicate the presence of requests or commitments within the paragraph. Because our annotated data is actually a collection of free text spans marked by multiple annotators, determining the correct label involves two steps: first we reconcile our annotations from multiple annotators into a single set of 'true' span annotations based on annotator agreement; secondly, we map the agreed annotations for a collection of spans into paragraph annotations. We explain each of these tasks in detail.

Measuring agreement between annotators for the span annotation task is significantly more complex than calculating message-level agreement, since selecting and marking text is not purely a text classification task. Importantly, the task does not have fixed units of annotation, as was the case when multiple annotators were labelling the same message units. For each span annotation, the human annotator is actually making at least two subjective decisions: which extent of text to annotate, and what label to apply to the text extent.

As discussed in Section 3.2.4, without fixed units of annotation, we cannot easily use traditional agreement measures of agreement such as Cohen's kappa, since we are not comparing independent labels being given to the same units of text. Instead, we rely on F-score, which is frequently used to calculate agreement in tasks where the unit of text being labelled is not fixed, e.g., (Hripcsak and Rothschild, 2005). To calculate the F-score, we measure the precision and recall between each pair of annotators. We treat the annotations of one annotator as the gold-standard and calculate the precision, recall and F-score for the other annotator's annotations against this reference.[3] For our task, we average the F-score across our pair of annotators. Across our 1000 messages,

[3] As already mentioned, it makes no difference to the F-score which annotator is selected as the gold-standard—swapping annotators swaps the values for precision and recall, but the F-score remains constant.

our two annotators identify 3397 spans with agreement of 0.86 for requests and 0.72 for commitments, measured by F-score.

5.1.2.1 Finding Agreed Spans

In our span annotation tool, text spans within a document are defined by three properties: the offset of the first character of the span relative to the message body (start offset), the span length, and the label annotated on the span.[4] In order to calculate the F-score between pairs of annotators, we require a metric for determining when two spans of marked text represent the 'same' utterance. We refer to these as metrics of span agreement.

The simplest possible span agreement metric is to require the start offset and span length to match exactly. Unfortunately, there are subtle effects of the annotation process that render such a measure unrealistic. For example, text spans marked by our annotators frequently differ only by the inclusion or exclusion of leading or trailing whitespace or utterance-final punctuation. We thus require a more nuanced metric to identify annotated spans that humans consider to mark the 'same' underlying utterance.

Before our span annotation task started, we ran a pilot annotation task to test our annotation tool and processes. This involved three annotators marking spans in 50 email messages. For this pilot task, exact span agreement was poor, with only 7 of 185 spans satisfying this metric. Adding just 1 character of tolerance at each end of the span, that is, allowing the start offset and end offset to differ by 1 character, resulted in an almost 10-fold increase in agreement, with 54 spans finding agreement. Human analysis of the additional matches was unanimous that all the spans should be considered the 'same'.

Figure 5.3 shows the number of agreed spans between three annotators using three different metrics with a variety of threshold values. In all cases, we normalise annotations for direction of annotation (left-to-right or right-to-left) before calculating agreement. The three metrics are:

1. Exact Match: Exactly equal values for start offset and span length;

2. Character Offset: Configured with a single threshold value that represents the maximum difference in both start offset and text length that is allowed for two spans to be considered the 'same'; and

3. Word Overlap: Configured with a threshold value that represents the minimum common word overlap required for two spans to be considered the 'same'.

As an example of how these metrics are applied, Figure 5.4 shows two different span annotations. Both have been labelled as a Request annotation. These would be assessed as follows:

[4] We also capture the text of the span and an optional free text comment that allows annotators to discuss issues of subjectivity or ambiguity for each annotation.

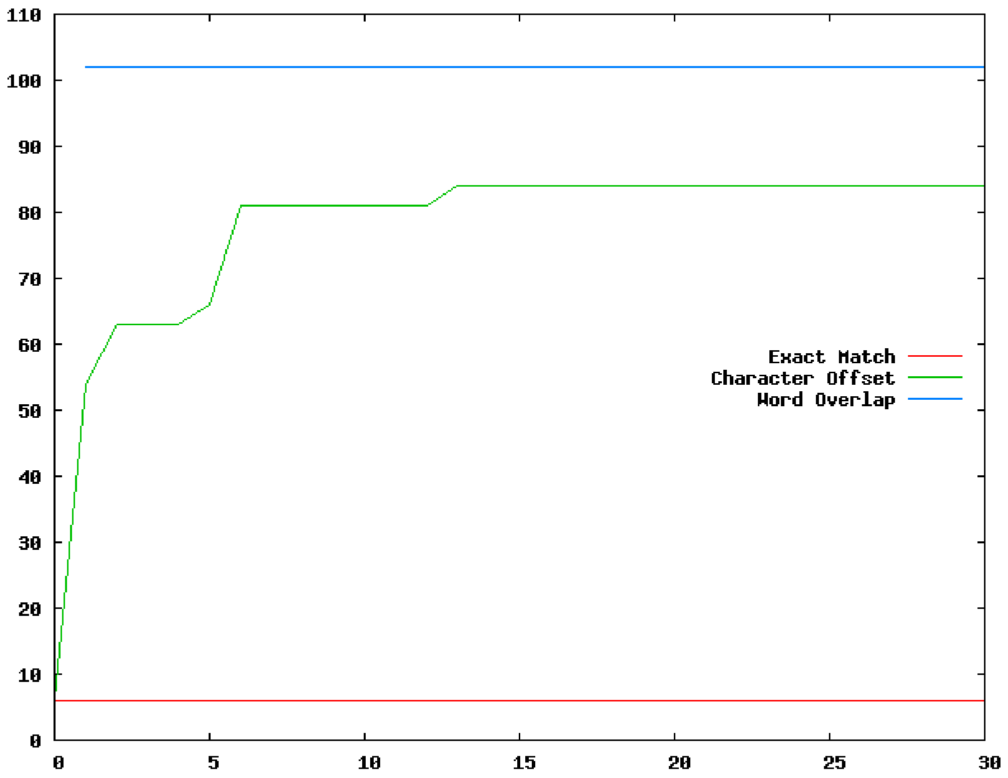
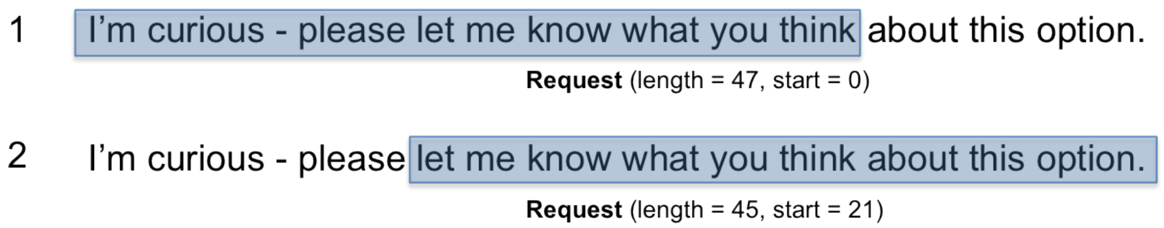


Figure 5.3: The absolute number of agreed spans, from a corpus marked with 185 spans, using each of three agreement metrics, with varying threshold values. The x-axis shows the value of the threshold variable; the y-axis shows the number of agreed spans.

1. Exact Match: Not a match.
2. Character Offset: The length differs by two characters and the start offset by 21 characters. These would be considered the same span only when the threshold value was greater than or equal to 21 (i.e., the greater of the difference in length or start offset).
3. Word Overlap: The spans share a six word overlap. They would be considered the same span only for a threshold value less than or equal to six.

For the Word Overlap metric, we also factor in the length of each span. If one text span overlaps all the words in the other span, as in Figure 5.5, then the two spans are considered the same, regardless of the overlap threshold. Elaborating on this example, the spans in Figure 5.5 would be assessed as follows under each of the metrics:

1. Exact Match: Not a match.
2. Character Offset: The length differs by 44 characters and the start offset differs by 7 characters. These would be considered the same span only when the threshold value was greater than or equal to 44.



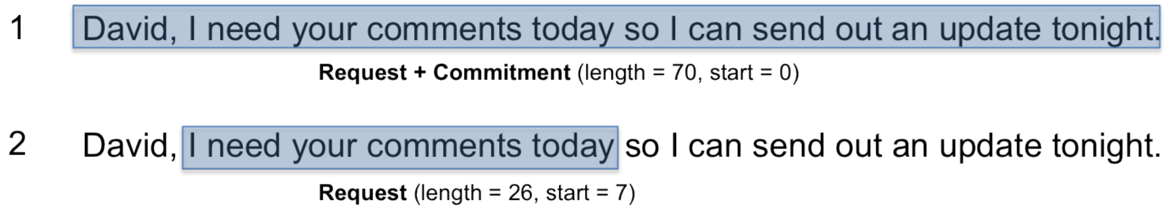
1 I'm curious - please let me know what you think about this option.

Request (length = 47, start = 0)

2 I'm curious - please let me know what you think about this option.

Request (length = 45, start = 21)

Figure 5.4: Two span annotations that are considered equivalent under our a word overlap metric with a threshold value of 6 or less.



1 David, I need your comments today so I can send out an update tonight.

Request + Commitment (length = 70, start = 0)

2 David, I need your comments today so I can send out an update tonight.

Request (length = 26, start = 7)

Figure 5.5: Two span annotations that are considered equivalent under a word overlap metric with any threshold value, since all the words in the second span overlap with the first. These are considered to be different spans under both the exact match metric and the character offset metric with a threshold value less than 44.

3. Word Overlap: The two spans have a five word overlap, which also represents the entirety of span 2. They are considered the same span for any threshold value.

As our two examples demonstrate, Exact Match is too restrictive. Figure 5.3 highlights this across our pilot task annotations, with only 7 of 185 annotated spans finding exact agreement matching spans. Measured with Character Offset with a two-character threshold[5], the number of span agreements increases by an order of magnitude, highlighting the frequent small differences that occur when marking the 'same' utterance. Further increasing the threshold value keeps increasing the number of matches, but for our data, this plateaus at a 15-character offset. In our dataset, this suggests that there are a collection of spans where some annotators have included a small number of extra words or punctuation characters in the span. As we saw in Figure 5.5, however, the start offset metric will not consider spans that mark a small substring of a larger span to be the same, unless the threshold value is increased to an unusually large number. Increasing the threshold to such a value is also likely to consider spans that mark disjoint and genuinely different spans to be the same, for example, annotations that mark separate short spans in sequential sentences. The

[5] Note that we use a single threshold value that captures the maximum allowable difference in both length and start offset for two spans to be considered matching.

plateauing in agreement numbers suggests that there is a reasonably wide gap between the threshold values required to identify the 'nearly the same' annotations, and annotations that mark small subsets of larger span annotations by another annotator.

From both the examples and the pilot annotation results, the Word Overlap metric is the most interesting. The metric is very stable, showing a complete lack of sensitivity to its threshold value, across our range of 1 to 30 word overlaps. This demonstrates that, while theoretically possible, we do not see any spans which overlap by even a single word that represent different spans.

Based on these results, throughout this chapter, we use the Word Overlap metric with its threshold set to 1 word to measure span agreement. A review of the literature reveals that this single token overlap is commonly used in evaluating agreement for named entity recognition— e.g., (Franzén et al., 2002; Tsai et al., 2006).

5.1.2.2 Calculating Gold-Standard Span Labels from Multiple Annotations

Having settled on the Word Overlap metric for matching spans, we then need to consider how to match span labels, remembering that our annotation scheme includes two dimensions: requests and commitments. We use a label matching metric that enforces agreement on the obligation act of interest (e.g., request), but ignores any disagreements for the other obligation act (e.g., commitment). Consider, for example, the two spans in Figure 5.5. Under our word overlap metric, these are considered to be the same span of text. Annotator A marked it as 'Request and Commitment', and Annotator B as 'Request'. If we were considering an exact label match, these two annotations would represent a disagreement. Under our looser label matching method, however, we consider these two span annotations to represent an agreed request annotation on the span, and a disagreed commitment annotation for the same span.

5.1.2.3 Generating Labelled Paragraphs

We use two different agreement metrics to aggregate our set of agreed annotated spans:[6]

1. Unanimous Agreement: only spans that are marked by both annotators with an equivalent label are included as gold-standard positive paragraphs. Spans where our two annotators disagreed are not considered to represent either an agreed positive or an agreed negative annotation, and are discarded; and

2. Superset Agreement: any span marked by either or both annotators as a positive span is included as a gold-standard positive paragraph. All other spans are included and considered to be negative paragraphs.

As we discuss further in Section 5.2.3, we use both these agreement metrics to generate training and test corpora to evaluate classifier performance. We convert a collection of agreed span annotations to paragraph annotations by considering any paragraph that

[6] Note that we do not consider a majority agreement metric, since we only have two annotators.

contains an agreed positive span annotation to have a positive annotation. All other paragraphs, except those that have been discarded due to disagreement when using the unanimous agreement metric, are labelled as negative paragraphs.

We refer to our derived corpora, including our paragraph corpora, as either **unanimous** or **superset** corpora, depending on the agreement metric used to generate them. Thus, when we speak of a 'unanimous request paragraph corpus' later in this chapter, we mean our corpus of paragraphs annotated for requests that contains only spans that are agreed by both annotators.

5.1.3 Generating a Gold-Standard Labelled Sentence Corpus from Annotated Spans

The process for creating a corpus of labelled sentences from our annotated span data is very similar. First, we segment and select relevant sentences, ignoring those that occur in irrelevant email zones. We then identify the correct gold-standard label for each selected sentence.

Before we can generate labelled sentences, however, we should at least briefly grapple with the notion of what constitutes a sentence. There are two main schools of thought towards defining sentences as linguistic units. The first of these considers a sentence as a grammatical unit, in which words are grammatically linked. Alternatively, a more form-based definition can be proffered, that considers a sentence from an orthographic perspective, where it is defined by anything contained between an initial capital letter and a full stop, question mark or other terminal punctuation marking. More than both of these, however, a sentence should also be internally coherent, meaning that the concepts and relations that underlie the surface text of the sentence, are linked, relevant and used to achieve the writer or speaker's communicative intent. For purely practical reasons of computational tractability, we approach sentences from the orthographic perspective for this experimental work, relying on the syntactic cues to distinguish boundaries between sentences so that they may be identified as units for classification.

To generate a corpus of labelled sentences, we first segment the message body text into sentences. Sentences generally do not span paragraph boundaries, so we build on the work already done for paragraph segmentation by starting from the message content that has already been segmented into paragraphs.

We split the text of each relevant paragraph into sentences, using the sentence splitter from the OpenNLP package (Apache Foundation, 2013), augmented with additional rules to handle common email sentence boundaries not seen in the news text on which the sentence splitter is trained. We classify email zones at the level of paragraphs, and consider sentences to have the same zone annotation as the paragraph from which they are extracted. We thus consider all sentences from a relevantly-zoned paragraph to also be relevant.

5.1.3.1 Calculating Gold-Standard Sentence Labels from Span Annotations

After segmenting the sentences and discarding those from irrelevant zones, we generate gold-standard binary request and commitment labels for each sentence. As for our paragraph corpus, because our annotated data is actually a collection of free text spans annotated by multiple annotators, determining the correct sentence label requires two steps: first we reconcile our annotations from multiple annotators into a single set of 'true' span annotations, and secondly we map the agreed annotations for a collection of spans into sentence annotations.

Again, we reuse the work done for paragraph corpus generation to identify the agreed spans, using the word overlap metric with a threshold value of 1. We also use the same metric to calculate agreed binary labels for each span.

Finally, we convert the span annotations to sentence annotations, using the same unanimous and superset agreement metrics outlined in Section 5.1.2.3. Any sentence that contains or overlaps with a positively-annotated, agreed span is given a positive annotation. There are three main classes of overlap:

1. **Span less than a sentence:** where an agreed span marks only a phrase or other part of a sentence. In this case, the sentence containing the agreed span is given the relevant label from the agreed span;

2. **Span equivalent to a sentence:** where an agreed span marks the full extent of a sentence. In this case, the sentence is given the relevant label from the agreed span; and

3. **Span more than a sentence:** where an agreed span includes content from more than one sentence. In this case, each sentence covered is given the relevant label from the agreed span, generating more than one positively labelled sentence.

All other spans, except those that have been discarded due to disagreement under the unanimous metric, are included in the sentence corpus as negative sentences.

5.2 Our Fine-grained Classification Approach

In this section, we describe our approach to paragraph and sentence classification, including how we build and evaluate our paragraph and sentence classifiers, and the techniques we employ to combat the increased imbalance that characterises some of our labelled corpora.

5.2.1 Classification Framework

Our approach to paragraph and sentence classification is built on a text-classification framework, using Support Vector Machine (SVM) classifiers with a Gaussian Radial Basis Function (RBF) as the kernel for the learning algorithm. Our selection of the RBF kernel is based on cross-validation experiments that demonstrated a small but consistent improvement in performance for our text classification tasks over an SVM employing a polynomial kernel. These results are consistent with Joachims' early experiments and published work on the use of SVMs for text classification—e.g., (Joachims, 1998).

For paragraph and sentence classification, we apply a range of mostly intra-paragraph or intra-sentence features to predict the correct binary label for whether the paragraph or sentence contains one or more requests and/or commitments. The exact features used differ for requests and commitments and are described in Sections 5.3 and 5.4. Most of the features are either lexical or based on shallow syntactic analysis.

We also pre-process the message text of each email message with our email zone classifier, Zebra, before classification. This ensures that our classifiers only consider content from relevant email zones when calculating features for classifying paragraphs and sentences. This is consistent with how we generate the gold-standard corpora that are used to train our classifiers, as described in Section 5.1.

5.2.2 Dealing with Skewed Data

A significant factor that affects the performance of our classifiers is that our fine-grained training and evaluation corpora are all skewed to varying extents. This occurs where one annotation label is significantly more frequent than other classes in the training data. We see particularly highly skewed data for commitments, with only 4% to 11% of paragraphs and sentences containing a commitment. Our data for requests is less imbalanced, with approximately 20–30% of paragraphs and sentences containing a request in our labelled corpora. Even these ratios are, however, significantly more skewed than our message-level request corpus from Chapter 4.

Highly imbalanced datasets, such as our commitment corpora, can cause significant challenges for automated classifiers. In this section, we describe how we adapt our approach to combat high levels of data skew in the cases where this occurs. High levels of imbalance also affect how we evaluate classifier performance. We are most interested in performance on paragraphs or sentences that actually contain requests and commitments. Ultimately, these are the paragraphs and sentences that we wish to identify. These positive instances are also the minority class for both request and commitment classification. Weighted F-score, a common metric for overall classifier performance, can be heavily dominated by performance on majority class instances under these circumstances. To combat this, we evaluate our classifier performance on both weighted F-score, which takes into account performance across all data, and on positive class F-score, which explicitly measures performance on instances that contain requests or commitments.

We experiment with three common approaches to mitigating the effects of data skew in our fine-grained corpora, which our review of related literature suggests are some of the most common techniques—e.g., (Schölkopf and Smola, 2002; Akbani, Kwek, and Japkowicz, 2004). These techniques are:

1. **Under-sampling** of the majority (negative) class;
2. **Over-sampling** of the minority (positive) class; and

3. **Cost-sensitive classification** where we specify variable costs for particular classes of errors. This can guide a classifier to compensate for data skew.

Under-sampling involves sampling a subset of instances that are labelled with the majority class. This reduces the skew of data seen by the classifier for training, creating a more balanced dataset, at the expense of discarding some labelled data from the majority class. The sampling process may be implemented in a number of ways, ranging from purely random selection of instances, through to ensuring that a representative sample of instances remains, with reference to the vector space of features that apply to each instance. For our experiments, we use the SpreadSubsample filter in Weka to randomly sample instances from our datasets while maintaining a maximum ratio between our minority and majority classes that is specified by a configurable threshold value.

Over-sampling is the complimentary technique to under-sampling: it attempts to balance the dataset by repeatedly sampling or synthesising additional instances that are labelled with the minority class label. At its simplest, this can involve duplicating existing minority class instances to balance the frequency of class instances. More usually, however, instances are synthesised by interpolating between existing instances with reference to the chosen feature space. We employ a popular variant of over-sampling called Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002).

Finally, cost-sensitive classification involves weighting different types of errors differentially, to bias the classifier towards or away from certain classes of error. In the case of an imbalanced dataset, this can be a useful technique to boost the relative importance of errors made on the minority class, despite their relative insignificance in the face of majority class errors. Consider, for example, a dataset of 1000 email messages of which 95% do not contain a commitment. Using accuracy as our metric, increasing the accuracy on messages without commitments by 10% would result in a 9.5% boost in accuracy. The same improvement in accuracy for messages that do contain commitments would boost overall accuracy by only 0.5%. In our case, however, that improvement in performance on commitment messages may be much more meaningful. Cost-sensitive classification allows us to bias the classifier towards or away from particular types of errors. We apply it to increase the cost of positive class errors, to counter the imbalance we see in some of our annotated corpora.

Each of these methods has its own advantages and drawbacks. Under-sampling, for example, discards valuable training data that is labour intensive for people to annotate. Under-sampling can mislead the classifier about likely ratios of positive and negative instances in unseen data. Over-sampling does not discard any data, but can have the same effect of misleading the classifier about likely class ratios ‘in the wild’. SMOTE, which synthesises additional minority-class instances between neighbouring actual positive instances, also suffers from this limitation. Cost-sensitive classification allows us to vary the relative weight or cost associated with different classes of error, namely false positive or false negative classification predictions. Varying the cost associated with these error classes allows us to bias the classifier towards either higher recall or higher precision.

We experiment with all three methods in building and evaluating standalone classifiers to identify requests or commitments. Where threshold values are required, such as the maximum ratio of the majority to the minority class for under-sampling and over-sampling, we use a standard GRID SEARCH approach to systematically evaluate a range of threshold values and pick the best performing values. Grid search is the *de facto* standard way of performing parameter optimisation. It simply involves an exhaustive search through a manually specified subset of the parameter space of the target parameter. A grid search algorithm must be guided by some performance metric, typically measured by cross validation on the training set or evaluation on a held-out validation set. In our case, we use cross validation over the same data to guide the parameter selection for the maximum class ratios in under-sampling and over-sampling.

Unfortunately, as our results in the next sections demonstrate, these techniques produced a disappointing lack of consistency in performance improvements. Despite this, however, one extremely important use of techniques like under-sampling and cost-sensitive classification is that they allow us to fairly directly control the trade-off between recall and precision in our classifiers. Later in this chapter, we leverage this capability to create high recall component classifiers when building the classification ensembles described in Section 5.5.

5.2.3 Evaluating Fine-grained Classification

One of the challenges in evaluating the performance of our fine-grained request and commitment classifiers stems from the number of parameters that are available to vary in training and evaluating each classifier.

Recall that our training and evaluation data for both our paragraph and sentence classifiers is drawn from a span-annotated corpus. In inducing labelled paragraph and labelled sentence corpora from these annotated spans, there are several variables that need to be considered around how we generate the training and evaluation data, and the variants of our classifiers that run over that data, including:

- the choice of unanimous or superset metrics for identifying agreed spans that are then mapped to our set of labelled training paragraphs;
- the tuning of constant and kernel parameters for our Support Vector Machine; and
- the tuning of parameters for under-sampling and cost-sensitive classification.

We tackle these dimensions of variation in a number of ways. Firstly, we evaluate the use of both unanimous and superset metrics for identifying agreed spans, as described in Section 5.1.2.3. Secondly, we experiment with feature selection to guide the number of features we deploy in each classifier. We do not explore variations in the feature set across experiments that seek to compare different classifier configurations (e.g., to compare our commitment paragraph classifier with and without under-sampling), though the features obviously do vary across different granularities and between our standalone and ensemble classifiers. Universally, however, we apply an information gain metric to select the top 5000 features for each classifier. Experiments with varying this number from 500 to 10,000 features revealed low levels of sensitivity to the number of features in our evaluation results, making us comfortable with running our experiments with a fixed-size feature set.

Finally, we exhaustively search for optimal parameters for our SVMs and our skew-reducing filters by undertaking a grid search across a well-defined search space to find optimal values for these parameters in each circumstance. The results that we present in this chapter, unless otherwise noted, represent the outcome of such grid search processes.

5.3 Fine-grained Request Classification

Our fine-grained request classifiers are built within the framework described in Section 5.2, using a Support Vector Machine as the learning algorithm, and trained with different subsets of the labelled corpora derived from our corpus of annotated spans, as described in Section 5.1.

In this section, we focus on classifying requests at two levels of granularity: the paragraph level and the sentence level. At each level of granularity, we describe the features used for classification, and then evaluate and analyse the performance of a range of differently configured request classifiers.

5.3.1 Paragraph-level Request Classification

We begin our fine-grained request classification at the paragraph-level. This level of granularity is useful to provide extracted summaries of actionable content in email messages, since paragraphs often provide enough context to be read and understood in isolation.

5.3.1.1 Paragraph-level Request Classification Features

We employ the following features for classifying requests in paragraphs:

- Lexical features in the form of binary word unigram and bigram features;
- Part-of-speech (PoS) information as binary PoS unigram and PoS bigram features;
- Paragraph length in words and characters;
- Presence of a recipient's name;
- Presence of the sender's name;
- Presence of a closed-set of words indicating urgency (e.g., *urgent*, *immediate*, *now*, *soon*);
- Presence of deadline feature, such as a day or date;

	Unanimous				Superset			
	# Inst	Baseline	SVM	Change	# Inst	Baseline	SVM	Change
Accuracy	3635	78.21%	91.58%	+17.09%	4025	70.63%	87.01%	+23.19%
Positive Precision	3635	0	0.900	-	4025	0	0.872	-
Positive Recall	3635	0	0.691	-	4025	0	0.653	-
Positive F-score	3635	0	0.781	-	4025	0	0.747	-
Weighted F-score	3635	0.782	0.912	+16.62%	4025	0.706	0.864	+22.38%

Table 5.1: Evaluation results for paragraph-level request classification.

- Presence of first- and second-person pronouns and possessives (e.g., *I*,*me*, *my*, *you*, *your*);
- Presence of question marks;
- Presence of modal verbs (e.g., *would*, *are*, *is*, *might*, *should*); and
- Presence of question words (e.g., *who*, *what*, *how*).

5.3.1.2 Paragraph-level Request Classification Evaluation

Table 5.1 shows the overall results for our paragraph-level request classifier. Results are compared against a simple majority class baseline. As with all majority class baselines for our fine-grained classifiers, the positive class precision, recall and F-score are all zero for the baseline system (since every instance is predicted to not contain a request). For comparing the performance of our SVM-based classifier against this baseline, we thus focus on weighted F-score and accuracy.

The baseline accuracy is reasonably high, ranging between 70.6% and 78.2% depending on which agreement metric we use to generate our training set. This, in turn, means that between 21.2% and 29.4% of paragraphs are agreed to contain a request. While this makes for a less balanced dataset than we saw for our message-level request training set, where 52% of messages contained a request, our results presented later in this section demonstrate that this level of imbalance has a relatively modest impact on classifier performance.

Our request paragraph classifier significantly outperforms the baseline classifier in terms of weighted F-score and accuracy at p=0.01. We report accuracy of 91.58%, over a ZeroR baseline of 78.21% using our unanimously agreed training and cross-validation dataset, and accuracy of 87.01% over a baseline of 70.63% for our larger dataset of all positively-annotated paragraphs. These results represent an increase of 17% and 23% over the baseline accuracy, respectively. From an error reduction perspective, this corresponds to a more than 60% reduction using our unanimously agreed dataset, and more than 55% for our superset dataset. In terms of weighted F-score, our classifiers outperform the baseline by between 16% and 22%.

	Unanimous			Superset		
	Existing	Under-sampling	Cost-Sensitive	Existing	Under-sampling	Cost-Sensitive
Accuracy	91.58%	91.97%	91.86%	87.01%	87.08%	86.71%
De-skewing Boost	-	+ 0.42%	+0.31%	-	+0.07%	-0.34%
Baseline Comparison	+17.09%	+17.59%	+17.45%	+23.19%	+23.29%	+22.77%
Positive Precision	0.900	0.858	0.886	0.872	0.847	0.836
De-skewing Boost	-	-4.67%	-1.56%	-	-2.87%	-4.13%
Positive Recall	0.691	0.756	0.718	0.653	0.684	0.681
De-skewing Boost	-	+9.40%	+3.91%	-	+4.75%	+4.29%
Positive F-score	0.781	0.804	0.794	0.747	0.757	0.751
De-skewing Boost	-	+2.94%	+1.66%	-	+1.34%	+0.54%
Weighted F-score	0.912	0.918	0.916	0.864	0.866	0.863
De-skewing Boost	-	+0.66%	+0.44%	-	+0.23%	-0.12%
Baseline Comparison	+16.62%	+17.39%	+17.14%	+22.38%	+22.66%	+22.24%

Table 5.2: Paragraph-level request classifier results, comparing the effect of under-sampling and cost-sensitive classification using unanimously-agreed and superset-of-positive-instance training datasets.

The data yielded by this evaluation provides strong evidence that we can reliably detect paragraphs that contain requests. This is encouraging, and builds on the strong results we achieved in Chapter 4 in identifying requests at the message-level.

We also quantify the headroom for further improving accuracy with techniques to reduce or compensate for the higher levels of data imbalance. As outlined above in Section 5.2.2, we experiment with three techniques: over-sampling our minority class (the positive request paragraphs), under- sampling our majority class (the non-request paragraphs), and configuring different costs for positive- class and negative-class errors via a cost-sensitive classification framework (making errors for positive request paragraphs more costly).

The headline results, generated via a grid search from a range of experiments conducted with these techniques, are shown in Table 5.2.

Table 5.2 compares the performance of our initial request paragraph classifier against the same classifier deployed with under-sampling and cost-sensitive classification filters. Performance is measured using accuracy, positive class precision, recall and F-score, and weighted F-score. We also compare performance against the original baseline classifier, to provide context to the metrics reported for our various de-skewed classifier configurations.

Readers may note that no results are included for over-sampling. Our experiments using grid search to optimise the parameters for Synthetic Minority Oversampling Technique (SMOTE) took multiple days to run on our available hardware, and produced results that were universally worse than our baseline classifier in terms of accuracy, weighted F-score and positive class F-score. As such, we halted our evaluation of this technique and did not generate an evaluation of SMOTE.

The optimal threshold values for the ratio between our positive and negative classes for under-sampling were calculated based on an exhaustive set of experiments that applied a grid search technique. We applied grid search using F-score as our performance metric, and applied the search across ten iterations of ten-fold cross validation for each set of threshold values and then averaged the results.

The results of our grid search demonstrated that the biggest boost to performance comes from under-sampling at a majority to minority class ratio of between 1.7 and 1.8 negative paragraphs for each request paragraph. For our classifier trained using unanimously agreed annotated data, applying under-sampling with a ratio of 1.7:1 boosted overall weighted F-score performance by almost 17.4%. For our superset of positively annotated request paragraphs, a ratio of 1.8:1 boosted weighted F-score by 22.66%. Accuracy results were increased by similar amounts. Positive class F-score—i.e., the F-score for paragraphs that contain requests—is an important measure of performance for such a skewed dataset. It was increased by between 1.34% and 2.94% using the under-sampling approach.

The other de-skewing approach we experimented with was cost-sensitive classification, which weights the cost of errors differently for each class. In our case, we want to weight errors for positive class (i.e., errors that are made with paragraphs actually containing requests) higher than errors that are made for paragraphs that do not contain requests. This is to counter the natural tendency of the algorithm to discount such errors, since they have a smaller effect on the overall accuracy due to the data skew. Because we focus on improving the positive class F-score, however, we want to avoid positive class errors as much as possible, while balancing this against the overall classifier accuracy and weighted F-score. As for under-sampling, we derived an optimal ratio of error costs for each class based on an exhaustive set of experiments run using a grid search technique. The optimal ratio of the relative cost of errors for the non-requests *versus* requests varied between 26% and 40% across our unanimous and superset training sets. These values led to a very small increase in weighted F-score of 0.44% for our unanimous dataset, and a drop in weighted F-score of 0.12% for our superset dataset, neither of which were statistically significant from our standalone classifier without cost-sensitive classification applied.

Overall, the performance increase gained from applying de-skewing techniques to our paragraph-level request classifier were small. Given the relatively modest levels of data skew observed, and the reasonably high performance of our paragraph-level request classifier in the face of this data skew, this is not surprising. While impossible to directly compare across different datasets, our paragraph-level weighted F-score of 0.912 is above even the F-score reported for many message-level request classification systems in the literature, including those evaluated in (Cohen, Carvalho, and Mitchell, 2004; Bennett and Carbonell, 2005; Carvalho and Cohen, 2006; Goldstein and Sabin, 2006). Based on the data from our evaluation, future work to further improve the performance of paragraph-level request classification would be best spent exploring factors other than the reduction of data imbalance.

	Unanimous				Superset			
	# Inst	Baseline	SVM	Relative Increase	# Inst	Baseline	SVM	Relative Increase
Accuracy	5882	78.60%	89.24%	+13.54%	6530	70.80%	85.05%	+20.13%
Positive Precision	5882	0	0.912	-	6530	0	0.862	-
Positive Recall	5882	0	0.550	-	6530	0	0.582	-
Positive F-score	5882	0	0.686	-	6530	0	0.694	-
Weighted F-score	5882	0.692	0.882	+27.46%	6530	0.581	0.837	+44.06%

Table 5.3: Evaluation results for sentence-level request classification.

5.3.2 Sentence-level Request Classification

Sentence-level request classification allows us to look towards services that might extract tasks from a user's inbox and reformulate them into a structured task list for ongoing management. In this section, we develop and evaluate classifiers for identifying sentences that contain requests.

The features that we apply for sentence-level request classification are the same as those we used for paragraph-level classification. We do add, however, some positional constraints on features we used for paragraph classification. These features include whether a sentence begins with a modal verb or ends with a question mark. As we do not attempt to segment paragraphs into sentences for paragraph classification, these features are not available at the paragraph level.

Table 5.3 presents the results for sentence-level request classification. The headline result when using our superset training set is that we are able to increase the accuracy over a majority baseline system up to 20% and improve the weighted F-score by more than 44%. This represents a highly significant 49% error reduction compared against our majority baseline. Again, all these results are generated using 10-fold stratified cross validation over the annotated data. For our classifier trained with our unanimous request training corpus, we see an increase in accuracy of 13.54% and an increase in weighted F-score of 27.46%, both of which are significant at $p = 0.01$.

The imbalance in our sentence corpus in terms of requests is very similar to the levels seen at the paragraph level: between 21.4% and 29.2% of sentences contain requests across our unanimous and superset training corpora. The lack of significant improvements we observed at the paragraph level, coupled with the similarities in features and levels of imbalance in our dataset, led us to not exhaustively evaluate the range of parameter settings for under-sampling and cost-sensitive classification techniques that we performed at the paragraph-level.

Using under-sampling with a maximum class ratio of 1.5, for example, we were able to boost positive class F-score by 2.9% to 0.706, and marginally increase weighted F-score from 0.882 to 0.883. Similarly, using cost-sensitive classification, we could increase positive class F-score by 3.9% to 0.713, and weighted F-score to 0.888. This increase in weighted F-score is, however, not significant, representing an increase of only 0.68% over our unfiltered SVM classifier.

Due to this lack of significant improvement when applying the various techniques for combating data skew, coupled with the volatility of the results—small changes in the threshold value resulted in relatively large changes in overall performance—we do not report the full set of results. We also do not consider that results using these techniques are likely to generalise to other datasets. As noted for our paragraph classifiers, these results suggest that there are factors other than data skew that would need to be focused on to further improve performance of our request sentence classifiers.

5.4 Fine-grained Commitment Classification

Like the request classifiers, our fine-grained commitment classifiers are built as described in Section 5.2, using Support Vector Machines trained with different subsets of the labelled corpora that are derived from our corpus of annotated spans, as described in Section 5.1.

As we have discussed in earlier chapters, one challenge in commitment classification is that there are weaker links between the surface form of an utterance and a commitment speech act; we do not see the relatively high frequency of a small number of forms as we do with imperatives and interrogatives for requests, for example.

Additionally, as our message-level classification experiments revealed in Chapter 4, commitments occur substantially less frequently than requests. This difference is even more pronounced at finer levels of granularity, which poses some additional challenges for classification.

We work to address these and other challenges in this section and in Section 5.5, where we look at building classifier ensembles to help counter the additional data skew seen at finer granularities.

5.4.1 Paragraph-level Commitment Classification

As already noted, our datasets of commitments are distinctly more imbalanced than those of requests, with the exact level of imbalance varying between our various different permutations of training corpora. At the paragraph level, only 4.2% of paragraphs are deemed to contain a commitment in our unanimous corpus; for our superset agreement corpus, the percentage is slightly higher, with 10.5% of paragraphs deemed to contain a commitment. Both these observed levels, however, are highly imbalanced, and are much higher than those we observe for requests.

This acute imbalance makes classification challenging. Firstly, there is very little headroom to outperform the high baseline that such data skew naturally dictates. Secondly, as discussed in Section 5.2.2, even large improvements to classification performance for the minority class (which we care most about) will have a relatively minor effect on overall classifier performance. These issues are demonstrated by the results we present in Section 5.4.1.2. We work to address them both later in this section, where we apply dataset rebalancing filters similar to those we used for paragraph classification in Section 5.3.1.

	Unanimous				Superset			
	# Inst	Baseline	SVM	Change	# Inst	Baseline	SVM	Change
Accuracy	3759	95.85%	97.77%	+2.00%	4025	89.52%	93.84%	+4.83%
Positive Precision	3759	0	0.909	-	4025	0	0.857	-
Positive Recall	3759	0	0.513	-	4025	0	0.495	-
Positive F-score	3759	0	0.656	-	4025	0	0.628	-
Weighted F-score	3759	0.938	0.975	+3.94%	4025	0.846	0.931	+10.05%

Table 5.4: Evaluation results for paragraph-level commitment classification.

5.4.1.1 Paragraph-level Commitment Classification Features

We employ the following features for classifying commitments in paragraphs:

- Lexical features in the form of binary word unigram and bigram features;
- Part-of-speech (PoS) information as binary PoS unigram and PoS bigram features;
- Paragraph length in words and characters;
- Presence of a recipient's name;
- Presence of the sender's name;
- Presence of words indicating urgency (e.g., *urgent*, *immediate*, *now*, *soon*);
- Presence of deadline feature, such as a day or date;
- Presence of first-person future tense phrases (e.g., *I will*, *I won't*, *we'd be*, *we would*);
- Presence of singular or plural first-person pronouns and possessives (e.g., *I*,*me*, *my*, *you*, *your*);
- Presence of continuing action with gerund verb form (e.g., *am doing*, *are doing*, *is doing*); and
- Presence of specified first person continuing phrases (e.g., *I am*, *we are*).

Some of these features, such as the first-person future tense phrases, are clearly linked to common forms of expressing commitments to future actions, such as *We'd be delighted to attend* or *I will send you the document tomorrow.*

	Unanimous			Superset		
	Existing	Under-sampling	Cost-Sensitive	Existing	Under-sampling	Cost-Sensitive
Accuracy	97.77%	97.77%	97.82%	93.84%	93.69%	94.09%
De-skewing Boost	-	0.00%	+0.05%	-	-0.16%	+0.27%
Baseline Comparison	+2.00%	+2.00%	+2.05%	+4.83%	+4.66%	+5.11%
Positive Precision	0.909	0.840	0.902	0.857	0.746	0.815
De-skewing Boost	-	-7.59%	-0.77%	-	-12.95%	-4.90%
Positive Recall	0.513	0.571	0.532	0.495	0.604	0.564
De-skewing Boost	-	+5.80%	+3.70%	-	+22.02%	+13.94%
Positive F-score	0.656	0.679	0.669	0.628	0.668	0.667
De-skewing Boost	-	+3.51%	+1.98%	-	+6.37%	+6.21%
Weighted F-score	0.975	0.976	0.975	0.931	0.934	0.936
De-skewing Boost	-	+0.10%	0.00%	-	+0.32%	+0.54%
Baseline Comparison	+3.94%	+4.05%	+3.94%	+10.05%	+10.40%	+10.64%

Table 5.5: Paragraph-level commitment classifier results comparing the use of various de-skewing techniques.

5.4.1.2 Paragraph-Level Commitment Classification Evaluation

Table 5.4 presents the overall results for our standalone, paragraph-level commitment classifier, before we employ any techniques for combating data skew. Results are again compared against a majority class baseline.

The high level of data skew and correspondingly higher baseline accuracies of 95.85% and 89.52% for our unanimous and superset classifiers is immediately noticeable. Despite the high baseline, our standalone commitment classifier manages to outperform it on both accuracy and weighted F-score, for both our superset and unanimous training and evaluation sets. These results are calculated using 10-fold stratified cross validation over the training data.

In situations such as this, where there is very little headroom to increase accuracy and weighted F-score, Error Reduction is often used as a metric for evaluating performance. For our unanimously trained classifier, we see an error reduction of 46.27%, and for our superset trained classifier, we see a reduction of 58.78%. These results provide strong evidence that our commitment classifiers credibly outperform their respective baseline classifiers.

Given the extreme imbalance in the dataset, however, we also hypothesise that deskewing the data for classifier training should provide a further boost to performance. We thus turn to the same range of skew-reducing classification techniques that we applied to our request classifiers. The results of applying under-sampling and cost-sensitive classification filters to our commitment paragraph classifier are reported in Table 5.5.

We optimise the ratio of positive to negative classes for our under-sampling filters using the same grid search approach that we applied with our request paragraph classifier. The optimal ratios of non-commitment paragraphs to commitment paragraphs

	Unanimous			Superset		
	Headroom	Actual	Error Reduction	Headroom	Actual	Error Reduction
Under-sampling Accuracy	2.28%	0.00%	0.00%	6.56%	-0.16%	-2.44%
Under-sampling Weighted F	2.56%	0.10%	3.90%	7.41%	0.32%	4.32%
Cost-sensitive Accuracy	2.28%	0.05%	2.19%	6.56%	0.27%	4.12%
Cost-sensitive Weighted F	2.56%	0.00%	0.00%	7.41%	0.54%	7.29%

Table 5.6: An analysis of performance headroom for paragraph-level commitment classification using different methods for combating data skew.

vary from 3.5:1 for our superset training data to 6.5:1 for our unanimous training data. Using these ratios in our under-sampling filter results in a marginal improvement in weighted F-score of between 0.10% and 0.32%. Overall accuracy is either unaffected or drops slightly, by 0.16%. None of these differences are, however, statistically significant. We do see substantial improvements in the F-score for our positive class, compared against our standalone paragraph classifier, with gains of between 1.98% and 6.21% for our unanimous and superset classifiers respectively. These gains come from significant increases in positive class recall, at the expense of smaller drops in positive class precision. This is the same pattern of results that we observed when applying under-sampling for request paragraph classification. It reflects the intended effect of under-sampling, which is to cause more instances to be classified as commitments, some of which will inevitably be incorrect classifications.

When we move to cost-sensitive classification, the optimal values for the relative cost of errors for the non-requests *versus* requests varied from 6% to 9% across our unanimous and superset training sets. This effectively makes a false positive error incur 6–9% of the weighting of false negative errors. These weight values were again derived from an exhaustive grid search process across a wide range of parameter values. As shown in Table 5.5, using cost-sensitive classification, we are able to lift accuracy by between 0.05% to 0.27%. Weighted F-score remains unchanged or improves by 0.54%. Positive class F-score is lifted by between 1.98% and 6.21%. The increase in positive class F-score is particularly encouraging, since this was the main source of errors in our commitment paragraph classifier.

As for our initial commitment classifier, the high levels of skew can make it useful to analyse the comparison between our classifiers from the perspective of error reduction. We do this in Table 5.6, explicitly drawing attention to the headroom for performance improvement and the actual improvement observed. Under our unanimous dataset, even a classifier with perfect accuracy can only lift accuracy by 2.28% and weighted F-score by 2.56% over our initial standalone commitment classifier. For our superset dataset, the headroom for improvement is 6.56% for accuracy and 7.41% for weighted F-score. Taken in this context, our marginal improvements over the standalone baseline still represent up to 12% of the best possible improvement in accuracy and weighted F-score. Table 5.6 details the actual performance increases we achieved against these theoretical maximums, and shows that, in the best case, we reduce the error rate for

	Unanimous				Superset			
	# Inst	Baseline	SVM	Relative Increase	# Inst	Baseline	SVM	Change
Accuracy	6106	95.89%	97.00%	+1.16%	6530	89.66%	92.28%	+2.92%
Positive Precision	6106	0	0.833	-	6530	0	0. 913	-
Positive Recall	6106	0	0.339	-	6530	0	0.280	-
Positive F-score	6106	0	0.482	-	6530	0	0.429	-
Weighted F-score	6106	0.939	0.964	+2.66%	6530	0.848	0.904	+6.60%

Table 5.7: Evaluation results for sentence-level commitment classification.

the weighted F-score by more than 7% when using cost-sensitive classification with our superset training data.

Overall, our paragraph classifiers outperform the majority baseline. Despite our application of a range of techniques to combat the reduced frequency of commitments in our email datasets, however, we failed to significantly improve the performance of these classifiers through the use of any skew-reducing techniques. While it is difficult to definitively identify the reason for this, it may be that our feature set fails to capture the essence of commitments as effectively as our request classifiers do. As we discussed back in Section 3.1.3.3, commitments lack a consistent syntactic form, unlike requests, which makes our task of classification using mostly surface-level lexical and syntactic features a more difficult one. It may also be that further improvements require a different approach, an hypothesis that we return to in Section 5.5 when we explore the application of classifier ensembles to combat the data skew.

5.4.2 Sentence-level Commitment Classification

The features we apply for sentence-level commitment classification are the same as those used for paragraph-level classification. Unlike our request classifiers, we do not employ explicit positional features for either paragraph or sentence classification, since there are no obvious lexical phrases that are highly indicative of commitments when they occur in particular sentence or paragraph positions.

Table 5.7 shows the results of our sentence-level commitment classifier using both our unanimous and superset training data, compared against our majority class baseline. All results are based on 10-fold stratified cross validation. These results show that our classifier manages to improve accuracy over the baseline by between 1–3%. These results are primarily limited by the extremely skewed dataset and correspondingly high baseline classifier accuracies of 95.89% and 89.66% for our two training sets. These values are marginally higher than those observed for our commitment paragraph corpus, illustrating an even higher level of data skew at the sentence level.

Weighted F-score is increased by between 2.5% and 6.6% over the baseline. Despite these increases in accuracy and F-score being small, they represent a significant error reduction of between 27% and 36.8%. That such small changes in accuracy and F-score correspond with such large error reduction rates highlights the small headroom

	Unanimous			Superset		
	Existing	Under-sampling	Cost-Sensitive	Existing	Under-sampling	Cost-Sensitive
Accuracy	97.00%	96.56%	97.05%	92.28%	92.57%	92.86%
De-skewing Boost	-	-0.45%	+0.05%	-	+0.31%	+0.63%
Baseline Comparison	+1.16%	+0.70%	+1.21%	+2.92%	+3.25%	+3.57%
Positive Precision	0.833	0.634	0.820	0.913	0.750	0.843
De-skewing Boost	-	-23.89%	-1.56%	-	-17.85%	-7.67%
Positive Recall	0.339	0.386	0.363	0.280	0.422	0.381
De-skewing Boost	-	+13.86%	+7.08%	-	+50.71%	+36.07%
Positive F-score	0.482	0.480	0.503	0.429	0.540	0.524
De-skewing Boost	-	-0.41%	+4.36%	-	+25.87%	+22.14%
Weighted F-score	0.964	0.962	0.965	0.904	0.916	0.916
De-skewing Boost	-	-0.21%	0.00%	-	+1.33%	+1.33%
Baseline Comparison	+2.66%	+2.44%	+2.77%	+6.60%	+8.02%	+8.02%

Table 5.8: Sentence-level commitment classifier results comparing the use of various deskewing techniques.

for improvement given the extreme imbalance in the annotated data, as we observed for paragraph classification.

Again, because of the extreme imbalance in the data, we turn to under-sampling and cost-sensitive classification as techniques to combat the data skew. The results of these experiments are reported in Table 5.8. We performed an exhaustive grid search of weighting parameters for both cost-sensitive classification and under-sampling to achieve the results reported in Table 5.8. Despite the extreme data skew, we do not see a consistent significant improvement to performance. When different class ratio and error cost values are applied, the results are highly variable. We see, for example, that the best case found through our grid search for under-sampling with our unanimously trained classifier was a ratio of 5 non-commitment sentences for each commitment sentence. This optimal configuration, however, still resulted in a small (and not statistically significant) drop in weighted F-score of 0.21% and of positive class F-score of 0.41%. For our superset trained classifier, we saw improvements in both weighted F-score (up 1.33%) and positive class F-score (up 25.87%) for under-sampling, with the ratio selected by grid search being 3:1.

Such variation makes it difficult to extrapolate performance to other data sets or to generalise the results. It is clear that under-sampling and cost-sensitive classification can improve performance as measured by weighted F-score and positive class F-score, sometime significantly. We do see consistency in the improvement in positive class recall that we have observed across all our fine-grained classifiers when employing under-sampling and cost-sensitive classification. There are equally, however, cases where it appears to hinder rather than improve performance.

5.5 Improving Fine-Grained Classification with Ensembles

The performance improvements observed from our application of under-sampling, over-sampling and cost-sensitive classification were inconsistent. While we saw an encouraging and consistent improvement in positive class F-score, performance measured by weighted F-score was much more varied. In light of these results, in this section we explore an alternative classification approach to boost the performance of our paragraph and sentence classifiers. While we apply these techniques to both our request and commitment classifiers, our primary focus is on commitment classification, given the very high levels of data skew.

When considering our task of identifying actionable requests and commitments in email, one important commonality across all the work in this thesis is that our message, paragraph and sentence classifiers are all attempting to identify the same underlying obligation acts in a given email message. Although the three types of classifiers are abstracting the text at different levels of granularity, all three levels of classifier are attempting to identify the same underlying signal, that is, some specific request or commitment utterance that is conveyed. Given this commonality, we hypothesise that we can share information between our classifiers, despite their different levels of granularity, to improve performance. Specifically, we explore the potential for improving request and commitment classification at the paragraph and sentence level by using information from coarser-grained classifiers to guide or direct the finer-grained classifiers.

The motivation for constructing such classifier ensembles is an underlying hypothesis that there are cues for classification available at the coarser message-level granularity that are not readily available for classification at the paragraph or sentence level. Message-level features such as subject-line text, sender and recipient information and choice of signoff phrase, for example, can be useful cues in classifying a message. Similarly, there is some evidence that using *Thanks* in a message signoff is positively correlated with the message containing a request. Such features are, however, generally not available to be represented in the feature space for paragraph or sentence classifiers, given that they are features of the message rather than any specific paragraph or sentence.

Additionally, there is greater lexical redundancy in a message than in a paragraph or sentence. This means that there may be more request and commitment words in a message than in a paragraph or sentence, and thus more cues for a classifier to recognise. On average, our data suggests that positive request messages have somewhere between 1.3–1.6 positive paragraphs per positive message. For commitments, we see approximately 1.2–1.6 positive paragraphs per positive message. While the paragraph classifier has visibility of cues of only a single paragraph for classification, the message-level classifier clearly has more cues, on average, available to help make a correct classification decision. Finally, message-level classification data is less skewed than paragraph or sentence classification. A more balanced dataset can make classification easier for most if not all machine learning classifiers.

Based on the additional cues and classification properties from the message level, we expect that making use of this information should result in higher accuracy for coarse-granularity classification. The results across our standalone classifiers confirm this; performance, measured by F-score on the positive class, monotonically drops as the granularity of classification is made finer for both request and commitment classification. Our coarse-to-fine grained classifiers aim to exploit this performance gap between our message and paragraph classifiers to improve paragraph classification.

We implement and evaluate two methods: the first, described in Section 5.5.2, uses our message, paragraph and sentence classifiers in a filtered ensemble to remove messages and paragraphs that are considered to not contain requests or commitments before finer-grained classification; the second approach, described in Section 5.5.3, augments the feature space of our paragraph classifiers with message-level predictions and that of our sentence classifiers with both message- and paragraph-level predictions.

5.5.1 Related Coarse-to-Fine Classification Work

While our motivations are a little different than other work in the literature, the idea of building ensembles of classifiers has been experimented with across a number of domains. Coarse-to-fine classification techniques have been used with great success in specific applications in areas such as computer vision (e.g., face detection) and to a more limited degree in natural language processing (e.g., parsing, machine translation). A workshop at the Neural Information Processing Systems (NIPS) Conference in 2010 focused on drawing together work in this field.[7] One paper at this workshop involved the coarse-to-fine classification of email for the purposes of spam detection (Pujara and Getoor, 2010). While this problem has some aspects in common with our own work, there are also important distinctions. Pujara and Getoor (2010) are motivated by managing the expense of large-scale classification rather than improving performance of a finer-grained task, and the disjoint nature of features that they applied at the different levels of classification make the outcomes not directly comparable with our own work. Their work focused on classification at four levels, listed below from cheapest to most expensive in terms of their spam detection analysis:

1. Initial `HELO` SMTP (Simple Mail Transfer Protocol) Connection: Attempt to detect spam cheaply based on the IP address from which the mail originates;

2. `Mail From` SMTP Command: Attempt to detect spam based on the sending address, as reported during the second command issued in an SMTP transaction;

3. Mail Subject: Attempt to detect spam based on the content of the email subject line; and

4. Mail Body: Analyse the mail body as a last resort, to confirm or refute spam classification.

[7] Information about the workshop is available from http://learning.cis.upenn.edu/coarse2fine/.

A coarse-to-fine approach is very appealing for their problem domain; email is generally delivered to Mail Transfer Agents (MTAs) using the Simple Mail Transfer Protocol (SMTP). This protocol defines a conversation consisting of a well-ordered set of commands. The first data provided to the MTA is the IP address of the remote sender, which is available as soon as the sender connects with a `HELO` command that is used to initiate a mail delivery from another mail server. As the conversation continues, the sender will send the `MAIL FROM` command and provide an originating email address. The sender then provides one or more recipient addresses using the `RCPT TO` command. Finally, the sender enters the `DATA` state and sends the body and full header content of the message. The structure of this conversation lends itself well to coarse-to-fine processing. In our case, we do not have the same structured sequence of data with increasing detail available to us for classification, since we are working from a corpus of already-delivered mail. Our goal is also different: we do not wish to reject mail from acceptance but to assist users to prioritise where their attention is best spent within their stream of delivered messages; all the messages we deal with fall into the finest (most expensive) category above. Finally, although Pujara and Getoor focus on multiple levels of classification, they do not explicitly look at cascading classifications to improve performance, rather they are focused on discarding spam messages as early and quickly as possible.

Zhang and Chow (2011) have also explored coarse-to-fine techniques for text classification. Their focus is on plagiarism detection, and they use the finer-grained document modelling primarily to retrieve other documents that should be analysed for possible plagiarism, rather than rely on a whole-of-document representation that produces a bigger set of less-relevant documents to analyse. Again, while their motivation differs from ours, the technique they apply to classify documents based on independent information from classifying paragraphs clearly shares the classification of the same phenomena at multiple levels of granularity in common with our approach.

Finally, work by McDonald et al. (2007) examines the classification of sentiment at various levels of granularity within documents. Their domain shares more in common with our own, in that the document-level sentiment classification should be reflected in the sentiment of finer-grained sentences. They employ a joint-structured model that factors the multiple levels of classification into a single model, rather than a cascaded system of classifiers such as ours. A joint model approach is something that we aim to consider in future work. McDonald et al. report encouraging results, including that document-level sentiment classification improved their sentence-level classification accuracy from 62.6% to 70.3%.

Ferguson et al. (2009) report similar findings exploiting paragraph-level classification results to improve document-level sentiment analysis of blog posts. Our work in this section aims to replicate and improve upon these results in the email domain, though from the opposite perspective, that is, exploiting document-level obligation act detection to improve finer-grained paragraph and sentence classification.

There is also other, more peripherally related, work on structured prediction—e.g., (Sapp, Toshev, and Taskar, 2010; Pedersoli, Vedaldi, and González, 2011)—and coarse-to-fine inference—e.g., Petrov's work in machine translation and parsing (Petrov and Klein, 2007; Petrov, Haghighi, and Klein, 2008; Petrov, 2012). Work in both these areas

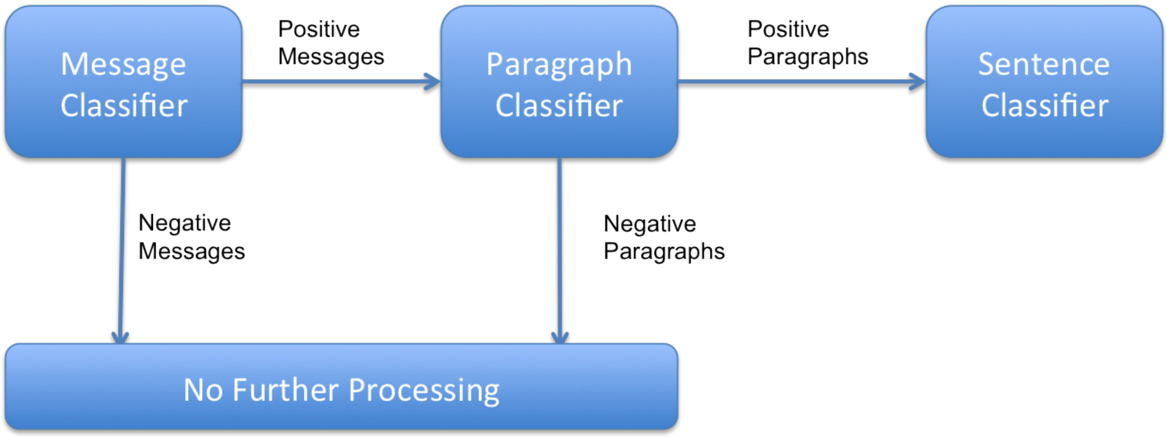


Figure 5.6: The filtered classifier ensemble, showing how messages are processed through a pipeline of message-level, paragraph-level and sentence-level classifiers, with negative instances being discarded at each processing point.

shares some aspects in common with our work, however, the focus on coarse-to-fine approaches tends to be as a method of making refined approximations due to intractable computational complexity, at the expense of classification accuracy/correctness. In contrast, we focus on improving fine-grained classification accuracy, rather than on optimising complexity or run-time performance.

5.5.2 Filtered Classifier Ensemble

Our first ensemble approach involves filtering input through a coarse-to-fine pipeline, as illustrated in Figure 5.6. We start by classifying each incoming email at the message-level, discarding messages that our message-level classifier is confident do not contain requests or commitments. The implicit assumption here is that messages that do not convey actionable intent will not contain any actionable paragraphs. We pass the remaining messages to our paragraph classifier for processing. This message-level filtering allows us to reduce the classification workload at the paragraph level, and more importantly, to reduce the skew of data for our paragraph classifier by ignoring all paragraphs in those messages that our message-level classifier is confident does not contain actionable content.

We repeat this process for sentence classification, filtering out paragraphs that are predicted to not contain requests or commitments by our paragraph classifier, and passing the remaining paragraphs to our sentence classifier for processing. This reduces the data skew for our sentence classifier even further, and also reduces the number of sentences that need to be processed.

False negative errors unavoidably propagate through such an ensemble; paragraphs and sentences in messages that are mistakenly marked as not containing requests or commitments will *never* be processed and thus can never be correctly identified by our paragraph and sentence classifiers. Because of this, we tune the coarser-grained classifiers to bias towards higher recall at the cost of lower precision to reduce this class of errors for fine-grained classification. We achieve this tuning using the techniques we applied earlier for de-skewing our data, namely cost-sensitive classification and under-sampling. This is equivalent to applying a confidence filter to our classifiers: only messages or paragraphs that our classifier is confident contain no actionable content of interest should be discarded.

5.5.2.1 Evaluating Filtered Classifier Ensembles

In all the classification results presented thus far, we have evaluated our classifiers using 10-fold, stratified cross validation, a widely accepted methodology for measuring the performance of a text classifier based on averaging its performance over slices of a single set of data that is repeatedly sampled for training and test data. We wish to evaluate the performance of our fine-grained classifiers under different ensemble configurations both intrinsically and extrinsically. The extrinsic evaluation is complicated by the way our coarse-grained classifiers in our filtered ensemble modify the data being passed to our paragraph and sentence classifiers at runtime. While incredibly useful for measuring performance against a baseline system, evaluation based on cross validation does not allow comparison across classifiers where the underlying dataset differs, as the results are effectively generated from different test sets. In the case of our classifier ensembles, because we filter out some instances, the set of data over which we evaluate our fine-grained classifiers in our Filtered Classifier Ensemble effectively changes each time different classification decisions are made at the coarse-grained level.

Given that we cannot use cross validation for evaluation, we instead create and use a single, held-out test set of annotated email data that is used for evaluation across all permutations of our ensemble classifiers. This test set is created from a separate set of 205 email messages that were annotated at the text span level by two independent annotators, using the same guidelines and annotation tool described in Section 5.1 to produce the training data. We transform this raw span-annotated test data into a consistent message-level, paragraph-level and sentence-level test set, using the same processes described in Section 5.1. As with our training data, we have to make a decision about the agreement we apply to generate our test sets. In order to provide a dataset that is as free of annotation noise as possible, we include only the unanimously agreed subset of annotated data, which makes the test set smaller than the original 205 messages, with the resulting test set size depending on the obligation act and granularity in question.

As well as using a separate test set, we also need to consider that our filtered ensemble offers two different ways in which we can measure classifier performance: intrinsic evaluation and extrinsic evaluation. In this context, we use intrinsic evaluation to refer to the performance of the classifier on the data which it actually processes. In our filtered classifier ensemble, intrinsic evaluation refers to measuring the performance of our paragraph classifier on the data that it processes, ignoring any data that was discarded by the message-level classifier. Extrinsic evaluation refers to

		Filter Ensemble			
	Standalone	Intrinsic	% Diff	Extrinsic	% Diff
Positive Precision	0.750	0.765	+2.00%	0.765	+2.00%
Positive Recall	0.761	0.815	+7.10%	0.638	-16.16%
Positive F-score	0.755	0.789	+4.50%	0.696	-7.81%
Weighted F-score	0.806	0.822	+1.99%	0.765	-5.09%

Table 5.9: Paragraph-level filtered request classifier ensemble performance compared with a standalone request paragraph classifier.

the overall performance of the paragraph classifier at a system level, including the implicit negative classifications that are given to any messages that are discarded by the message classifier.

5.5.2.2 Filtered Request Classifier Ensemble

As our request classifiers have less data skew to deal with than our commitment classifiers, there is less potential for our classifier ensembles to assist. The remaining hypotheses about coarser-grained classifiers having access to stronger and more redundant signals, however, still applies. It is for this reason that we build and evaluate a filtered request classifier ensemble.

We first evaluate the performance of our filtered request classifier ensemble at the paragraph level. Table 5.9 shows both the intrinsic and extrinsic evaluation of our filtered request classifier ensemble at the paragraph level, measured against our separate, held-out test set of gold-standard paragraphs, derived from our separate annotated corpus of 205 email messages.

Readers will note that the results for our standalone request paragraph classifier, trained using our superset annotated paragraph corpus, differ from the results presented in Section 5.3.1.2. This is because those results were generated using 10-fold cross validation, while the results in Table 5.9 are calculated over our separate test set. All the results in this section are calculated using the separate test set to allow direct comparison of the results.

The performance of our classifier ensemble, as measured through the intrinsic evaluation results, is uniformly better than our standalone classifier, with weighted F-score increasing by 1.99% and positive F-score by 4.50%. When measured extrinsically, however, our ensemble classifier shows a modest drop in both positive class F-score and weighted F-score of 7.81% and 5.09% respectively.

A closer look at the data, through the lens of a confusion matrix, confirms the source of classification errors. Table 5.10 clearly shows that false negative errors are the largest class of errors. Of the fifty false negative paragraph classification errors, only 40% are actually errors from the paragraph classifier; 60% of errors result are due to errors

	Intrinsic Results		Extrinsic Results	
	Actual True	Actual False	Actual True	Actual False
Predicted True	88	27	88	27
Predicted False	20	172	50	297

Table 5.10: The confusion matrix for our paragraph-level filtered request classifier ensemble.

		Filter Ensemble			
	Standalone	Intrinsic	% Diff	Extrinsic	% Diff
Positive Precision	0.641	0.752	+17.32%	0.752	+17.32%
Positive Recall	0.641	0.783	+22.15%	0.516	-19.50%
Positive F-score	0.641	0.767	+19.66%	0.612	-4.52%
Weighted F-score	0.725	0.769	+6.07%	0.714	-1.52%

Table 5.11: Sentence-level filtered request classifier ensemble performance compared with a standalone request sentence classifier.

made by the message classifier which the paragraph classifier has no chance to correct or circumvent. This error analysis led us to explore alternate classifier ensembles that allow our finer-grained classifiers the chance to recover from coarse-grained classification errors where there is sufficient fine-grained evidence to discount the coarse-grained decision. We describe and evaluate this second approach in Section 5.5.3.

At the sentence level, we are also able to improve the intrinsic performance of our standalone request sentence classifier, as shown in Table 5.11. Positive class F-score is increased by almost 20%, and weighted F-score by more than 6%. Despite this, however, we again see a modest drop of 1.5% in extrinsically measured weighted F-score, and a corresponding decrease in positive class F-score of 4.52%.

Error analysis again reveals that false negative errors are the largest class of error, and that 70% of the extrinsic false negative errors stem from false negative paragraph classification that are actually cascaded errors made by our paragraph and message classifiers. It is also important to note that these results are generated after we have tuned the recall of our message-level classifier to reduce false negative errors and to maximise the weighted F-score for our paragraph and sentence classifiers through cross validation.

5.5.2.3 Filtered Commitment Classifier Ensemble

As for requests, we developed an ensemble of filtered classifiers for commitments. The configuration is the same, with our message, paragraph and sentence classifiers combined into a pipeline, as shown in Figure 5.6.

	Intrinsic Results		Extrinsic Results	
	Actual True	Actual False	Actual True	Actual False
Predicted True	112	37	112	37
Predicted False	31	115	105	572

Table 5.12: The confusion matrix for our sentence-level filtered request classifier ensemble.

		Filter Ensemble			
	Standalone	Intrinsic	% Diff	Extrinsic	% Diff
Positive Precision	0.507	0.696	+37.28%	0.696	+37.28%
Positive Recall	0.717	0.762	+ 6.28%	0.604	-15.76%
Positive F-score	0.594	0.727	+22.39%	0.646	+8.75%
Weighted F-score	0.717	0.798	+11.30%	0.759	+5.86%

Table 5.13: Paragraph-level filtered commitment classifier ensemble performance compared with a standalone commitment paragraph classifier.

We first evaluate performance at the paragraph level, with results summarised in Table 5.13. The data from this evaluation lends support to our hypothesis that ensemble classification was likely to provide more boost to commitment classification than for requests. We see performance of our commitment paragraph classifier boosted across all intrinsic metrics, and more importantly, we see extrinsically measured positive class F-score and weighted F-score increase by 8.75% and 5.86% respectively. We see a huge boost to positive class precision of more than 37%.

Our error analysis at the paragraph level, summarised in the confusion table in Table 5.14, again reveals false negatives to be the largest source of errors under extrinsic evaluation, with 52% of these errors propagating from message-level classification errors.

To explore the headroom for further performance improvement in our filtered classifier, we also tested it with a 'perfect' message-level classifier, one that would not propagate any errors to the paragraph classifier. We do this by using gold-standard message-level annotations to calculate feature values for finer-grained classifiers, rather than relying on the output of our actual message-level classifier. These results are summarised in Table 5.15. The headline result is that, measured extrinsically, the same paragraph classifier can improve its own performance by more than 8% on weighted F-score and by more than 14% on positive class F-score when coupled with a message classifier that makes no message-level classification mistakes. This validates our hypothesis that a filtered classification ensemble approach has a strong potential to improve fine-grained classification accuracy.

	Intrinsic Results		Extrinsic Results	
	Actual True	Actual False	Actual True	Actual False
Predicted True	32	14	32	14
Predicted False	10	141	21	374

Table 5.14: The confusion matrix for our paragraph-level filtered commitment classifier ensemble.

	Intrinsic			Extrinsic		
	Filter	Gold Msg	% Diff	Filter	Gold Msg	% Diff
Positive Precision	0.696	0.760	+9.20%	0.696	0.760	+9.20%
Positive Recall	0.762	0.717	-5.91%	0.604	0.717	+18.71%
Positive F-score	0.727	0.738	+1.51%	0.646	0.738	+14.24%
Weighted F-score	0.798	0.801	+0.38%	0.759	0.821	+8.17%

Table 5.15: Paragraph-level filtered commitment classifier ensemble performance with gold-standard message-level classification

A closer look at the results in Table 5.15 reveals the expected pattern that results for positive class precision, recall and F-score do not vary from intrinsic to extrinsic results for our ensemble with gold-standard message classification. We do, however, see a boost in weighted F-score, when the non-commitment messages that were correctly discarded by our message classifier are included in the extrinsic evaluation results. Interestingly, the positive class recall for our gold-standard ensemble is actually lower than that measured intrinsically in our baseline filter commitment paragraph ensemble. This result suggests that some of the commitments that are discarded at the message-level in our filter ensemble are genuinely difficult for our paragraph classifier to detect, even when they are not discarded at the message-level.

Moving to commitment sentence classification, we are interested to measure whether the positive results and potential we have seen at the paragraph level are repeated at the sentence level. The results of our filtered commitment classifier ensemble at the sentence level are summarised in Table 5.16.

Unfortunately, although our intrinsic sentence classifier results show an increase in positive class F-score of more than 30% and in weighted F-score of 11.5%, the extrinsic evaluation of both positive class and weighted F-score is weaker than our standalone commitment sentence classifier, showing a drop of 6.02% and 2.60% respectively.

Again, our error analysis reveals that the biggest source of error is an increase in false negatives, leading to a positive class recall for our filtered classifier that is 43% lower than the intrinsic results and 23.90% lower than that of our standalone classifier. As can be seen from the confusion matrix results in Table 5.17, more than 70% of these false negatives stem from the cascaded errors from the earlier pipeline classifiers.

		Filter Ensemble			
	Standalone	Intrinsic	% Diff	Extrinsic	% Diff
Positive Precision	0.575	0.729	+26.78%	0.729	+26.78%
Positive Recall	0.523	0.700	+33.84%	0.398	-23.90%
Positive F-score	0.548	0.714	+30.29%	0.515	-6.02%
Weighted F-score	0.693	0.773	+11.54%	0.675	-2.60%

Table 5.16: Sentence-level filtered commitment classifier ensemble performance compared with a standalone commitment sentence classifier.

	Intrinsic Results		Extrinsic Results	
	Actual True	Actual False	Actual True	Actual False
Predicted True	35	13	35	13
Predicted False	15	99	53	682

Table 5.17: The confusion matrix for our sentence-level filtered commitment classifier ensemble.

This result provides further evidence that our filtered classification ensemble approach struggles at finer granularities, due to cascaded errors. At the sentence level in particular, there are two levels of classification which cascade errors to the finer-granularity classifiers that are unrecoverable. This is the major constraint on classification performance at the sentence level, with more than 57% of all errors made at the sentence level stemming from cascaded errors in earlier stages of the filter ensemble. This can be clearly seen in the confusion matrix in Table 5.17.

Just as for our commitment ensemble at the paragraph level, we are interested to further quantify how much our sentence classification results are affected by cascaded errors, and to understand the headroom for improving sentence-classification performance through improvements in our coarser-grained classifiers. To achieve this, we re-evaluated our sentence-level classification results using 'perfect' message-level and paragraph-level classifiers, that is, classifiers that do not propagate any errors to the sentence classifier. These results are summarised in Table 5.18. The headline result is that, measured extrinsically, the same commitment sentence classifier can improve on its own performance by more than 17.7% on weighted F-score and by more than 35.5% on positive class F-score when coupled with message and paragraph classifiers that make no message-level or paragraph-level classification mistakes. Coupled with the evidence already seen at the paragraph level, this provides overwhelming evidence to corroborate our hypothesis that a filtered classification ensemble approach has a strong potential to improve fine-grained classification accuracy.

Even in the face of the encouraging results around the potential of our filtered

	Intrinsic			Extrinsic		
	Filter	Gold Msg	% Diff	Filter	Gold Msg	% Diff
Positive Precision	0.729	0.728	-0.14%	0.729	0.728	-0.14%
Positive Recall	0.700	0.670	-4.29%	0.398	0.670	+68.34%
Positive F-score	0.714	0.698	-2.24%	0.515	0.698	+35.53%
Weighted F-score	0.773	0.728	-5.82%	0.675	0.795	+17.78%

Table 5.18: Sentence-level filtered commitment classifier ensemble performance with gold-standard message-level and paragraph-level classification

classifier ensemble for sentence-level commitment classification, the fact remains that the real-world results of our filtered commitment classifier at the sentence level are the most disappointing results so far. From a pragmatic perspective, our filtered ensemble failed to outperform our standalone commitment sentence classifier.

The take-away message from our filtered classification ensemble experiments is that we see strong results and even stronger potential for commitment classification at the paragraph and sentence level using a filtered classifier ensemble. For requests, the results are less positive, with real-world performance showing a drop in positive class and weighted F-score for both paragraph and sentence classification. In the case of both requests and commitments, however, the data from our experiments strongly suggests reducing the number of cascaded false negative errors is the most pressing issue to address to further improve results across all four fine-grained classifiers.

In the next section, we address exactly this issue, developing and evaluating an alternate approach to our classifier ensembles that attempts to avoid cascaded errors from coarse-grained classifiers.

5.5.3 Feature-Augmented Classifier Ensemble

The results from our series of experiments in Section 5.5.2 clearly demonstrate that the primary weakness of our filtered classification ensemble approach stems from errors in coarse-grained classifiers that compound through the classifier levels to reduce the performance of finer-granularity classifiers. For paragraph and sentence classification, every false negative error made by our message classifier is unrecoverable; it is impossible for either paragraph or sentence classifiers to correctly identify positive obligation paragraphs or sentences in a message that our message classifier mistakenly marks as negative, because these are never passed to the finer-granularity classifiers for processing. As a result, finer-grained recall of the ensemble system is negatively affected to coarse-grained errors, often dramatically so.

To combat this limitation, we modify our classifier ensembles to embed coarse-grained classification predictions in the feature space of finer-grained classifiers, rather than using the coarse-grained classifiers to filter out negatively predicted instances. For

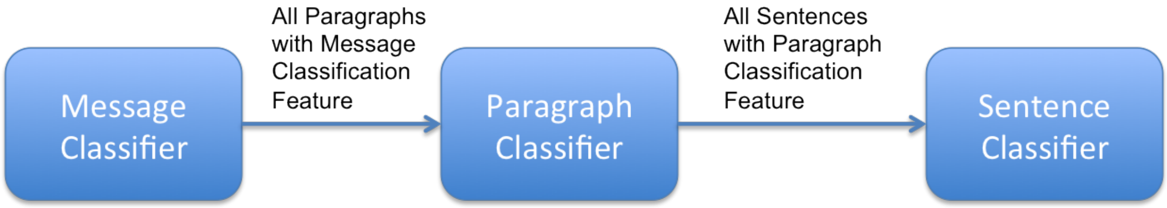

Message Classifier
All Paragraphs with Message Classification Feature
Paragraph Classifier
All Sentences with Paragraph Classification Feature
Sentence Classifier

Figure 5.7: The feature-augmented classifier ensemble, showing how messages are processed through a pipeline of message-level, paragraph-level and sentence-level classifiers, with features being added to the feature space at each processing point.

example, the message classifier adds its message classification prediction as a feature for each paragraph from that message. This means that our paragraph classifier has access to the predicted message classification for each paragraph, as an attribute in the feature space for each paragraph. The process is repeated for finer granularities, meaning our sentence classifier has access to both the predicted message classification and the predicted paragraph classification for the message and paragraph from which each sentence is drawn. We refer to this as a feature-augmented classifier ensemble, and the basic configuration is illustrated in Figure 5.7.

From an evaluation perspective, each classifier in our feature-augmented classifier ensemble processes every message or paragraph or sentence, so does not require the same distinction that we drew between intrinsic and extrinsic performance for our fine-grained classifiers in our filtered classifier ensembles.

Given the promising results we saw in Section 5.5.2 for boosting the performance of our fine-grained commitment classifiers, we focus solely on commitment classification in this section. Coupled with the disappointing performance in our filter classifiers, the lower levels of underlying data skew for requests strongly suggests that even our augmented ensemble classifiers will have less positive impact for request classification than for commitments. Of course, our request classifiers also already have very credible performance in a standalone configuration.

To further test the hypothesis that feature-augmented classifier ensembles do not provide significant assistance for request classification, we did briefly evaluate the performance of a feature-augmented request classifier ensemble. While we do not present the full results here, we did observe moderate improvements in weighted F-score over the filter ensemble performance of 1% or less at the paragraph and sentence level, suggesting that our feature-augmented classifier ensemble does go some way to correcting the propagation of false negative errors. Measured against the standalone request classifier performance, however, the differences were not statistically significant for either sentence or paragraph classification. We do not explore these results for request classification further, as the results indicate that standalone classification provides the best trade-off in terms of complexity and performance for request classification.

			Feature-Augmented	
	Standalone	Filter Ensemble	Extrinsic	% Improvement
Positive Precision	0.507	0.696	0.912	+31.03%
Positive Recall	0.717	0.604	0.585	-3.15%
Positive F-score	0.594	0.646	0.713	+10.37%
Weighted F-score	0.717	0.759	0.806	+6.19%

Table 5.19: Paragraph-level feature-augmented commitment classifier ensemble results compared against the equivalent filtered ensemble and standalone paragraph classification results.

5.5.3.1 Feature-Augmented Commitment Classifier Ensemble

We first focus on paragraph-level results for a feature-augmented commitment classifier ensemble. Table 5.19 shows the paragraph classification performance of our feature-augmented commitment classifier. The first column of Table 5.19 shows the results for our standalone commitment paragraph classifier. The second column shows results for the same classifier configured in a filtered classifier ensemble, as we presented in Section 5.5.2.3. The third column shows results for the same paragraph classifier, configured in our feature-augmented classifier ensemble. Finally, we show the relative performance improvement, compared against the filtered ensemble classifier. We measure the improvement against the filtered ensemble classifier instead of the standalone classifier, since this provides a higher baseline than the standalone classifier, and reflects the performance boost over our highest performing classifier configuration thus far.

As we see from the results, our feature-augmented classifier outperforms our standalone paragraph classifier by more than 12% and our filtered classifier ensemble by more than 6%, measured by weighted F-score. An additional positive result is that the feature-augmented classifier outperforms the filtered classifier by more than 10%, and the standalone classifier by more than 20% on positive class F-score. These represent the best performance figures attained for commitment classification over our held-out test set of all the approaches explored in this thesis.

Error analysis of these results shows that the feature-augmented classifier actually does not reduce false positive errors as expected. To the contrary, we see a very slight increase, with one additional false negative prediction. Interestingly, the paragraphs given a false negative classification are different from those that were incorrectly classified in the filtered classifier. This suggests that in retraining our paragraph classifier with the additional message-prediction feature, we have naturally repositioned the support vector in our SVM, and thus different classification decisions are being made for the same data. Given that we also see false positive prediction reduced by more than 78%, the data suggests that the lack of decrease in false negatives is a trade-off made by the classifier to achieve the much lower level of false positive classifications. On the

	Filtered Ensemble		Augmented Ensemble	
	Actual True	Actual False	Actual True	Actual False
Predicted True	32	14	31	3
Predicted False	21	374	22	385

Table 5.20: The confusion matrix for our paragraph-level feature-augmented commitment classifier ensemble contrasted against our filtered commitment classifier.

			Feature-Augmented	
	Standalone	Filter Ensemble	Extrinsic	% Improvement
Positive Precision	0.575	0.729	0.552	-4.00%
Positive Recall	0.523	0.398	0.602	+15.11%
Positive F-score	0.548	0.515	0.576	+5.11%
Weighted F-score	0.693	0.675	0.711	+2.60%

Table 5.21: Sentence-level feature-augmented commitment classifier ensemble results compared against the equivalent filtered ensemble and standalone sentence classification results.

basis of this data, our feature-augmented classifier does appear to address the weakness of our filtered classifier ensemble at the paragraph level, despite the overall false negative classification rate not being reduced. Additionally, our predicted message classifications do appear to help guide the paragraph classifier away from false positive predictions that it otherwise makes.

When we move onto commitment sentence classification results, the results are equally encouraging. As summarised in Table 5.21, the headline result is that our feature-augmented sentence classifier outperforms both our filtered classifier ensemble and our standalone commitment sentence classifier by 5.33% and 2.60% respectively, when measured by weighted F-score. Measured by positive class F-score, the results are even more positive, outperforming the standalone classifier by 5.11%, and the filtered classifier by 11.84%. These represent the highest level of performance in classifying commitment sentences across all the experiments in this thesis.

We perform an error analysis using the confusion matrix presented in Table 5.22. Unlike at the paragraph level, we see a significant decrease of almost 34% in false negative classifications. This data strongly supports our hypothesis that our feature-augmented ensemble should reduce the false negatives observed in our filtered ensemble. Coupled with this decrease in false negatives, we see a corresponding rise in false positives, but obviously this increase is low enough that both positive class F-score and weighted F-score are higher for sentence classification in our augmented ensemble.

In summary, the results of our feature-augmented classification ensembles confirm

	Filtered Ensemble		Augmented Ensemble	
	Actual True	Actual False	Actual True	Actual False
Predicted True	35	13	53	43
Predicted False	53	682	35	652

Table 5.22: The confusion matrix for our sentence-level feature-augmented commitment classifier ensemble contrasted against our filtered commitment classifier.

this as the most accurate approach for classifying commitments at the paragraph level and sentence level in the email data we have worked with in this thesis. This result confirms our hypothesis that leveraging coarser-grained classifiers that are built on data with less skew can provide useful signal to classifiers working to identify the same commitments phenomena at finer levels of granularity.

5.6 Summary

In this chapter, we have systematically developed and evaluated a range of fine-grained classification methods for identifying paragraphs and sentences that contains requests and commitments. As part of this activity, we have presented a series of approaches to deal with the data imbalance at these finer levels of granularity that particularly affects commitment classification.

We began by describing the process and outputs from our span-level annotation task that generated the fine-grained annotations required for training supervised machine learning classifiers at the paragraph and sentence level. Section 5.1 describes, in detail, how we determine agreement between annotators from the free text spans identified and labelled, and how we map these into labelled corpora of paragraphs and sentences for both our request and commitment classifiers.

Using datasets derived from this corpus, we train and evaluate several fine-grained request classifiers, which we present in Section 5.3. Our results clearly demonstrate that we can achieve very satisfactory performance at the paragraph level with a standard SVM classifier, achieving a weighted F-score of 0.912 in 10-fold cross validation. Additional work to apply under-sampling, over-sampling and cost-sensitive classification made, at best, a marginal difference to performance. Based on the data from our evaluation, we can conclude that future work to further improve the performance of paragraph-level request classification would be best spent exploring factors other than the reduction of data imbalance. Our request sentence classification results tell a similar story. Our SVM-based classifier was able to outperform our baseline system by between 27% and 44%, with a larger relative improvement being achieved when training and testing with a gold-standard dataset based on superset agreement between annotators. Largely, however, the difference in performance between superset and unanimous datasets can be ascribed to differences in majority baseline performance, with superset training and test sets leaving more headroom for improvement due to lower levels of data skew. We saw marginal, but not significant, improvements in applying cost-sensitive classification and under-sampling. Again, in cross validation, the classifier performs at a level that is high enough to deploy for real-world use, although we acknowledge that performance on unseen data will undoubtedly be lower.

For commitments, discussed in Section 5.4, the story was quite different. Our annotated commitment paragraph and sentence corpora are highly skewed: less than 5% of sentences and 10% of paragraphs contain a commitment. Given this skew, we again applied techniques such as under-sampling, over-sampling and cost-sensitive classification with the expectation that they should lead to performance improvements. We performed exhaustive grid search of a range of tuning parameters for these filters. Unfortunately, none of these techniques made a significant difference to our classification performance; sentence classifier performance was either significantly worse or not significantly different from unfiltered performance.

Despite the variability in the results of our attempts to tackle data skew within our request and commitment classifiers, one thing that is consistent across all our under-sampling and cost-sensitive classification experiments at both the paragraph and sentence level is that positive class recall is significantly increased, at the expense of lower positive class precision. In the context of the end applications of our work, where our goal is to identify all actionable content in email, this bias of increasing recall at the expense of precision is likely to be an acceptable and even desirable trade-off, as the cost of missing an obligation is higher than the cost of filtering through some spuriously identified actions. This makes our under-sampling and cost-sensitive classification techniques applicable for our problem domain, not only for specific cases where they can directly increase overall performance, but also for their ability to bias our classifiers towards increasing positive class recall.

The coarse-to-fine classifier ensembles that we describe and evaluate in Section 5.5 leverage exactly this capability to tune positive class recall. Their performance confirmed our hypothesis that additional information and lexical cues are available at coarser granularities and that this additional information can be successfully exploited to improve finer-granularity text classification. Our filtered classifier ensembles collectively demonstrate we can improve the intrinsic performance of paragraph and sentence classifiers so that they exceed the standalone classifier performance. Their performance is hampered, however, in terms of overall extrinsic performance measured by weighted F-score and positive class F-score, which often fails to meet or exceed standalone classifier performance. Error analysis clearly identifies that the most significant source of errors for all our filtered ensembles stems from false negative classifications compounding through the classification pipeline, errors that are caused by errors propagated from coarser-grained classifiers.

This insight led us to experiment with feature-augmented classifier ensembles that reliably improve on the performance of both filtered ensemble and standalone classifiers for commitment classification at both the paragraph and the sentence level. Thus the feature-augmented commitment classifier ensemble represents our highest performing fine-grained commitment classification mechanism.

Summarising the ground covered in this chapter, we have produced two sets of classifiers for reliably identifying requests and commitments at the fine-granularity required to enable the development applications, such as automated, task-focused summarisation and automated task extraction from email messages, that we envisage could help support task management in email. We begin exploring some of these applications in the prototype software that we present in Chapter 7. For requests, our fine-grained classifiers strongly resemble the approach we successfully applied for message classification in Chapter 4. For commitments, however, our novel coarse-to-fine classifier ensemble, built on augmenting the feature space of fine-grained classifiers with coarse-grained predictions, outperforms all the other approaches we evaluated. This is an important result that may well have application in areas beyond our own work. While the related work that we surveyed in Section 5.5.1 explores coarse-to-fine processing ideas to reduce the processing load while incurring the smallest possible drop in performance, our work pushes these ideas towards a different boundary. Our results suggest that coarse-to-fine classification approaches may well have a role to play in boosting the performance of fine-grained classifiers where data skew makes classification challenging.

"One of my pet-peeves is when someone does not get back to me, but I am one of the worst offenders."

Email user, quoted in (Whittaker and Sidner, 1996, p. 277)

6 Integrating Obligation Classifiers with Email Client Software

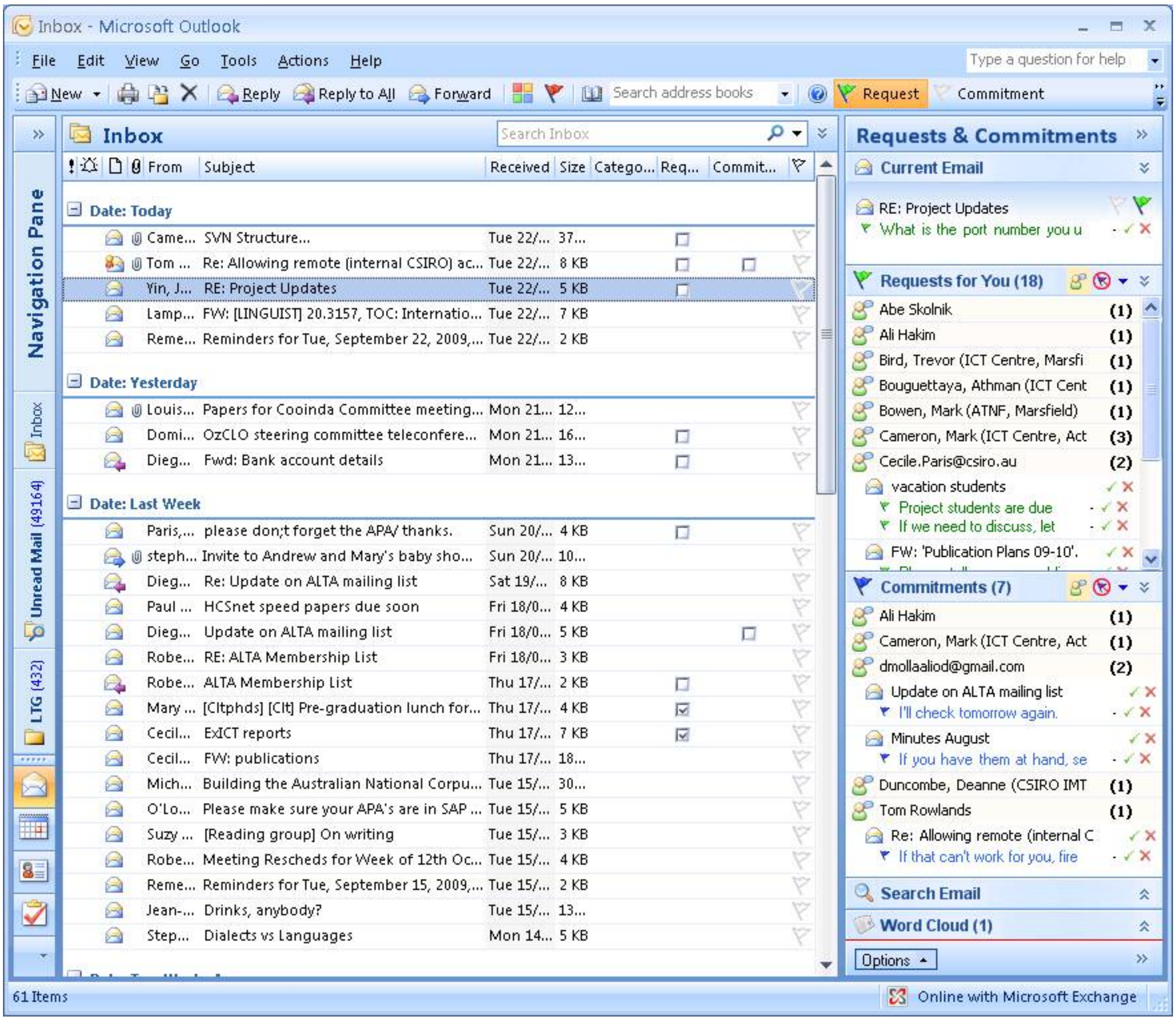
Having developed and intrinsically evaluated a series of request, commitment and email zoning classifiers in Chapters 4 and 5, in this chapter we present a software plug-in that we have designed and implemented which integrates our automatic request and commitment classifiers within Microsoft Outlook, the most widely-used business email habitat.[1] The plug-in creates a software platform that offers a range of features for identifying, displaying and interacting with obligation acts in email messages. We describe these features in Section 6.1.

To understand how the plug-in is being used, and how users are interacting with identified tasks, the plug-in is deeply instrumented, allowing the detailed recording of user actions and feedback. We present details of this instrumentation and logging functionality in Section 6.2. The data that is captured through this instrumentation and logging can be used to iteratively improve classifier performance, and can also enable use of the tool for in-context annotation.

In Section 6.3, we then discuss current and future applications of the plug-in. Primarily, the plug-in functions as an application for email task management, to improve task handling within a user's inbox. It does so by enabling requests and commitments to be both automatically and manually identified. It uses our request and commitment classifiers for automatic obligation act identification, and users can also manually identify additional obligation acts or correct automatically identified tasks.

We also envisage the plug-in being used for manual annotation, as a natural environment for manual, in-context annotation of speech acts or other pragmatic phenomena

[1] Statistics for email client software usage are regularly published at http://emailclientmarketshare.com/; last accessed here 27th May 2013.



Inbox

From	Subject	Received	Size	Req...	Commit...
Date: Today					
Came...	SVN Structure...	Tue 22/...	37...	□	
Tom ...	Re: Allowing remote (internal CSIRO) ac...	Tue 22/...	8 KB	□	□
Yin, J...	RE: Project Updates	Tue 22/...	5 KB	□	
Lamp...	FW: [LINGUIST] 20.3157, TOC: Internatio...	Tue 22/...	7 KB		
Reme...	Reminders for Tue, September 22, 2009,...	Tue 22/...	2 KB		
Date: Yesterday					
Louis...	Papers for Cooinda Committee meeting...	Mon 21...	12...		
Domi...	OzCLO steering committee teleconfere...	Mon 21...	16...	□	
Dieg...	Fwd: Bank account details	Mon 21...	13...	□	
Date: Last Week					
Paris,...	please don;t forget the APA/ thanks.	Sun 20/...	4 KB	□	
steph...	Invite to Andrew and Mary's baby sho...	Sun 20/...	10...		
Dieg...	Re: Update on ALTA mailing list	Sat 19/...	8 KB		
Paul ...	HCSnet speed papers due soon	Fri 18/0...	4 KB		
Dieg...	Update on ALTA mailing list	Fri 18/0...	5 KB		□
Robe...	RE: ALTA Membership List	Fri 18/0...	3 KB		
Robe...	ALTA Membership List	Thu 17/...	2 KB	□	
Mary ...	[Cltphds] [Clt] Pre-graduation lunch for...	Thu 17/...	4 KB	☑	
Cecil...	ExICT reports	Thu 17/...	7 KB	☑	
Cecil...	FW: publications	Thu 17/...	18...		
Mich...	Building the Australian National Corpu...	Tue 15/...	30...		
O'Lo...	Please make sure your APA's are in SAP ...	Tue 15/...	5 KB		
Suzy ...	[Reading group] On writing	Tue 15/...	3 KB		
Robe...	Meeting Rescheds for Week of 12th Oc...	Tue 15/...	4 KB		
Reme...	Reminders for Tue, September 15, 2009,...	Tue 15/...	2 KB		
Jean-...	Drinks, anybody?	Tue 15/...	13...		
Step...	Dialects vs Languages	Mon 14...	5 KB		

Requests & Commitments

Current Email

RE: Project Updates
What is the port number you u

Requests for You (18)

Abe Skolnik	(1)
Ali Hakim	(1)
Bird, Trevor (ICT Centre, Marsfi	(1)
Bouguettaya, Athman (ICT Cent	(1)
Bowen, Mark (ATNF, Marsfield)	(1)
Cameron, Mark (ICT Centre, Act	(3)
Cecile.Paris@csiro.au	(2)
vacation students	
Project students are due	
If we need to discuss, let	
FW: 'Publication Plans 09-10'.	

Commitments (7)

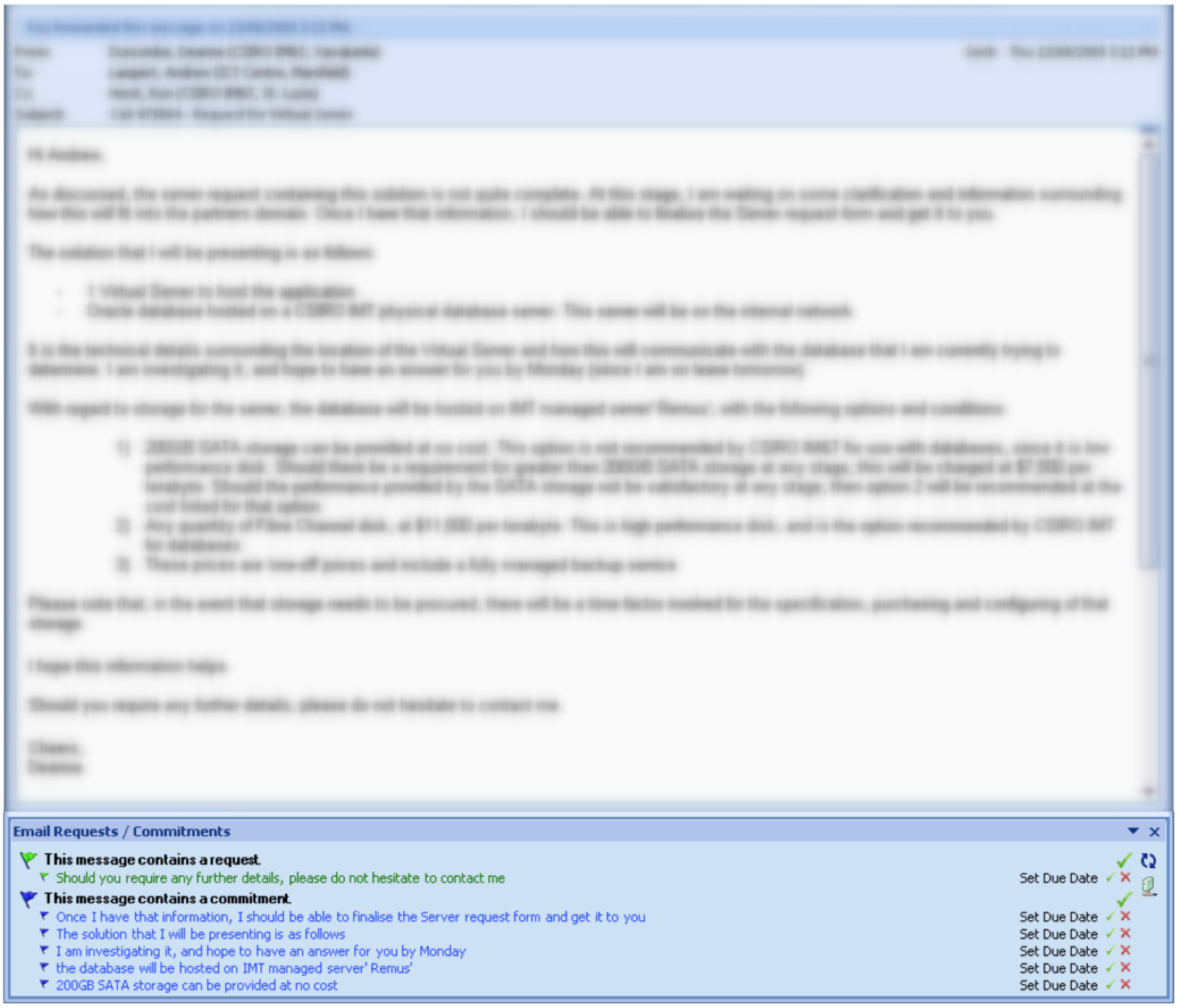
Ali Hakim	(1)
Cameron, Mark (ICT Centre, Act	(1)
dmollaaliod@gmail.com	(2)
Update on ALTA mailing list	
I'll check tomorrow again.	
Minutes August	
If you have them at hand, se	
Duncombe, Deanne (CSIRO IMT	(1)
Tom Rowlands	(1)
Re: Allowing remote (internal C	
If that can't work for you, fire	

Search Email

Word Cloud (1)

Figure 6.1: The Email plug-in shown within Microsoft Outlook.

in email messages. In this use, the instrumentation and logging allows for annotation judgements to be recorded for each message in which a user identifies obligation acts. Building on this capability, we have also considered the plug-in as a platform for extrinsic evaluation of automatic request and commitment classifiers. The plug-in is designed to support real-world email use to make both the annotation and evaluation functionality as ecologically valid as possible. It does so for evaluation by allowing the assessment of a range of automated classification and email processing tools within real users' email habitats. While we apply the plug-in to working with obligation content in email, it could also be readily adapted and used to identify, label and display other phenomena in email text, such as topics or sentiment. These extensions, along with a complete execution of an extrinsic user evaluation of our automated classifiers, are left to future work.



Email Requests / Commitments

This message contains a request

- Should you require any further details, please do not hesitate to contact me — Set Due Date ✓ ✗

This message contains a commitment

- Once I have that information, I should be able to finalise the Server request form and get it to you — Set Due Date ✓ ✗
- The solution that I will be presenting is as follows — Set Due Date ✓ ✗
- I am investigating it, and hope to have an answer for you by Monday — Set Due Date ✓ ✗
- the database will be hosted on IMT managed server' Remus' — Set Due Date ✓ ✗
- 200GB SATA storage can be provided at no cost — Set Due Date ✓ ✗

Figure 6.2: The Message View Panel, shown below the Outlook message window.

6.1 Features of the Outlook Plug-in

The overall appearance of the main plug-in interface is as shown in Figure 6.1.

6.1.1 Message View Panel

The **Message View Panel** (Figure 6.2)[2] is displayed at the bottom of the standard Outlook message window when an email is opened. The Message View Panel shows all the requests and commitments within the email message being viewed.

Accompanying the requests and commitments are a set of controls, similar to those in the Task Sidebar that we describe in Section 6.1.2, for the user to interact with

[2] Content is blurred for privacy.

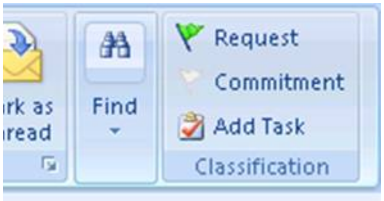


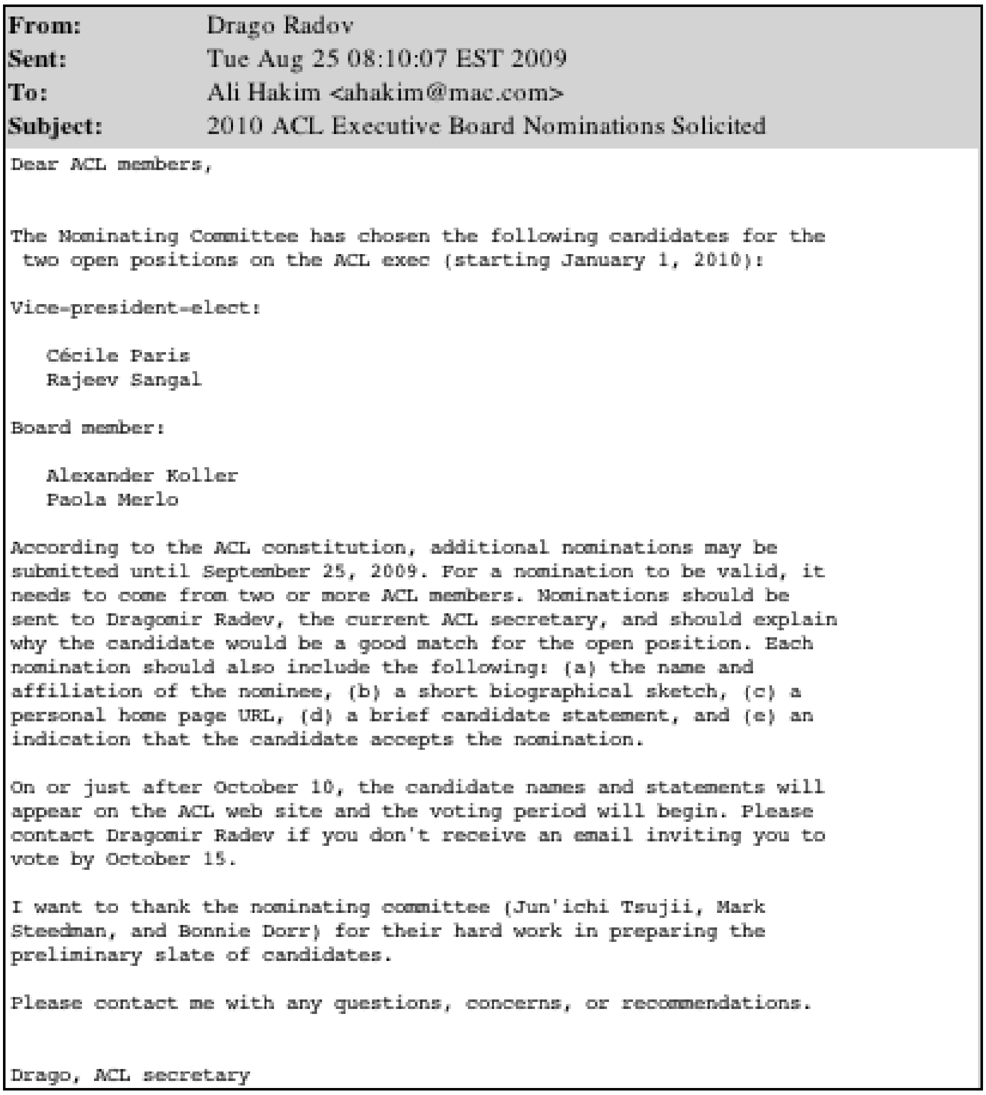
FIGURE 6.3: The Request and Commitment Buttons in the Microsoft Outlook ribbon. The coloured Request Flag and greyed-out Commitment Flag indicate that the message contains at least one request and no commitments.

and manage this set of tasks. For any request or commitment, the user may set a due date, mark it as complete, or delete it (if it has been identified in error, or if the user simply does not want to identify that request or commitment as a task to be tracked). Additionally, as described further in Section 6.2.3, the user is able to inspect and modify the current logging policy associated with the message. The logging policy is how the user controls whether the current email message will have its content and actions logged for the purposes of iteratively improving classifier performance, capturing annotations, or evaluating classifier performance. The user can opt-out from such logging for any specific message, or using rules that automatically exclude matching messages from being logged.

To support the Message View Panel, three buttons are added to the standard Outlook ribbon that is visible while an email message is being read: a Request Button, a Commitment Button and an Add Task Button, as shown in Figure 6.3. In Figure 6.3, the coloured Request Flag and greyed-out Commitment Flag indicate that the current message contains at least one request and no commitments.

When no message text is selected, clicking on either the Request or Commitment Button while reading a message annotates a message-level request or commitment for that message, without specifying any specific associated text span. To identify a specific text span as a request or commitment, the user highlights some text (e.g., a phrase or a sentence) and clicks on the request or commitment button. This identifies a fine-grained request or commitment, along with a message-level request or commitment if one has not already been marked. The Request and Commitment Buttons also function as status indicators, showing whether a request or commitment has been manually or automatically identified. These Request and Commitment Buttons also work while working in the inbox, while a user views a list of messages in the current folder or inbox, as shown in the left-hand pane of Figure 6.1. In this context, when clicked, the Request and Commitment Buttons create message-level requests or commitments for messages that are selected in the current folder. It is only possible to identify message-level requests and commitments in this way; fine-grained obligation acts can only be identified while viewing the content of a message.

Finally, the Add Task button also shown in Figure 6.3 allows the selected text to



From: Drago Radov
Sent: Tue Aug 25 08:10:07 EST 2009
To: Ali Hakim <ahakim@mac.com>
Subject: 2010 ACL Executive Board Nominations Solicited

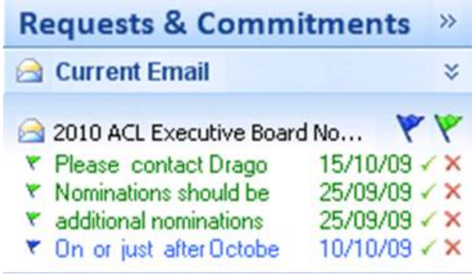
```
Dear ACL members,


The Nominating Committee has chosen the following candidates for the
 two open positions on the ACL exec (starting January 1, 2010):

Vice-president-elect:

   Cécile Paris
   Rajeev Sangal

Board member:

   Alexander Koller
   Paola Merlo

According to the ACL constitution, additional nominations may be
submitted until September 25, 2009. For a nomination to be valid, it
needs to come from two or more ACL members. Nominations should be
sent to Dragomir Radev, the current ACL secretary, and should explain
why the candidate would be a good match for the open position. Each
nomination should also include the following: (a) the name and
affiliation of the nominee, (b) a short biographical sketch, (c) a
personal home page URL, (d) a brief candidate statement, and (e) an
indication that the candidate accepts the nomination.

On or just after October 10, the candidate names and statements will
appear on the ACL web site and the voting period will begin. Please
contact Dragomir Radev if you don't receive an email inviting you to
vote by October 15.

I want to thank the nominating committee (Jun'ichi Tsujii, Mark
Steedman, and Bonnie Dorr) for their hard work in preparing the
preliminary slate of candidates.

Please contact me with any questions, concerns, or recommendations.


Drago, ACL secretary
```

Figure 6.4: An email message containing requests and commitments that are extracted and illustrated in the Current Email Pane in Figure 6.5.

be used to create a structured Outlook task that is managed separately from the user's email messages. This button was added on the basis of user feedback, but we have very little insight into how widely such a function would be used.

6.1.2 Task Sidebar

The **Task Sidebar**, shown on the right-hand panel of Figure 6.1, provides an overview of obligations from the user's mailbox. It contains five separate information panes. The top three panes are used to display and interact with the requests and commitments



Requests & Commitments »

Current Email

2010 ACL Executive Board No...

Please contact Drago	15/10/09	✓	✗
Nominations should be	25/09/09	✓	✗
additional nominations	25/09/09	✓	✗
On or just after Octobe	10/10/09	✓	✗

Figure 6.5: The Current Email Pane showing a task-focused summary for the email message in Figure 6.4. The email shown contains three requests (in green) and one commitment (in blue).

that have been either manually annotated or automatically identified in email messages in the user's inbox and other selected folders. The final two panes are the Search Pane, which allows the user to search through the identified requests and commitments, and a Word Cloud Pane that provides an overview of the frequent words in the requests and commitments identified.

We describe each of the five information panes below.

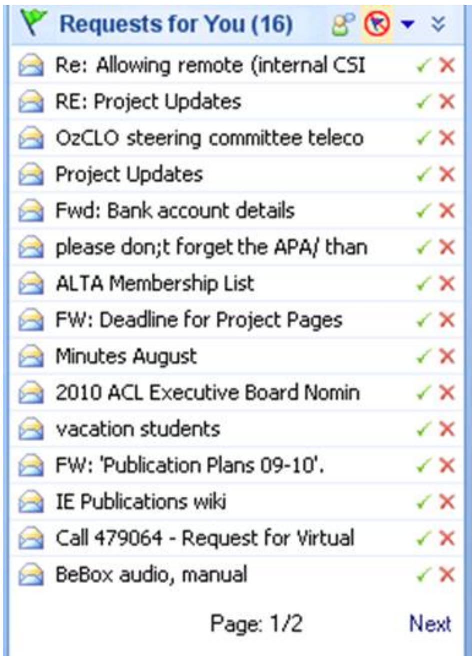
6.1.2.1 Current Email Pane

The **Current Email** pane, shown in Figure 6.5, contains two flags which indicate whether the currently selected message contains any requests and commitments,[3] along with a list of any specific text spans that have been identified as containing requests or commitments. The message in Figure 6.4 contains three requests and one commitment that have been identified and displayed in Figure 6.5. Each of these obligation has a due date specified, and none have been marked as complete.

The purpose of this pane is to provide a short, task-focused summary of the obligation content in a single message, to give the user a quick, browsable sense of the obligations to which they might need to respond or which they might wish to follow up. Within the Current Email pane, the user can interact to mark a request or commitment as complete, by clicking on the green tick next to each extracted obligation act. To delete a request or commitment that has been mistakenly identified, the user can click on the red cross next to each obligation act.

Finally, to set or modify the due date for each sentence-level task, the user can click on the current date. Sentence-level tasks with a due date show that date alongside, as in Figure 6.5. Clicking on this date allows the user to edit the due date. Message-level displays of tasks do not display dates or allow due dates to be set, as can be seen in Figure 6.6. Sentence-level tasks without an associated due date display a single – character in the place of a due date, as can be seen next to the each fine-grained task

[3] Here, green for requests and blue for commitments.



Requests for You (16)

Re: Allowing remote (internal CSI	✓ ✗
RE: Project Updates	✓ ✗
OzCLO steering committee teleco	✓ ✗
Project Updates	✓ ✗
Fwd: Bank account details	✓ ✗
please don;t forget the APA/ than	✓ ✗
ALTA Membership List	✓ ✗
FW: Deadline for Project Pages	✓ ✗
Minutes August	✓ ✗
2010 ACL Executive Board Nomin	✓ ✗
vacation students	✓ ✗
FW: 'Publication Plans 09-10'.	✓ ✗
IE Publications wiki	✓ ✗
Call 479064 - Request for Virtual	✓ ✗
BeBox audio, manual	✓ ✗

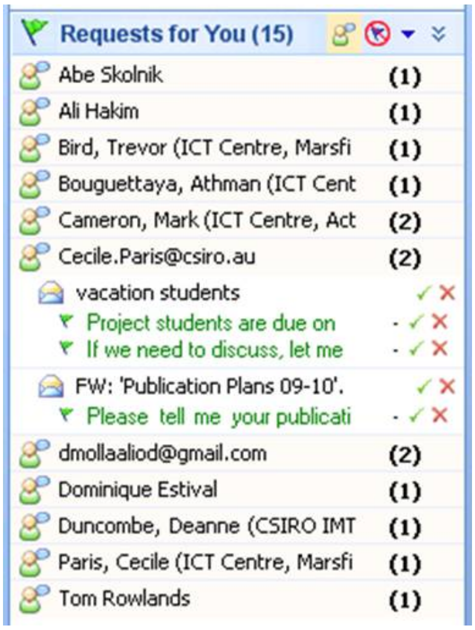
Figure 6.6: The Requests Pane showing tasks at the message level for a collection of email messages.

in Figure 6.7. This character can be clicked to associate a due date with each task.

In the current implementation, the Current Email pane shows the message-level classification, as well as a simple list of any request and commitment spans that have been identified. In future work, it should be possible to integrate functionality to automatically rewrite the text of extracted requests and commitments to ensure the obligations listed in this pane makes sense as standalone text, for example, by resolving pronouns and potentially combining multiple realisations of the same underlying request or commitment into a single obligation. This task rewriting functionality could also be applied when extracting requests and commitments to a separate, structured task list.

6.1.2.2 Requests Pane

The **Requests** Pane, as shown in Figure 6.6 and Figure 6.7, displays the outstanding messages with obligations that are awaiting action by the mailbox user. Requests can be shown at the message level, as in Figure 6.6, or can include detail of the specific request utterances within each message, as in Figure 6.7 and Figure 6.8. Where a message-level view is offered, the email subject is displayed to indicate that one or more obligations for the mailbox user are contained in that message. The finer-granularity displays allow for the display of more than one obligation act per email message, as is shown for some messages in Figure 6.8 and Figure 6.7.



Requests for You (15)

Abe Skolnik (1)

Ali Hakim (1)

Bird, Trevor (ICT Centre, Marsfi (1)

Bouguettaya, Athman (ICT Cent (1)

Cameron, Mark (ICT Centre, Act (2)

Cecile.Paris@csiro.au (2)

vacation students ✓ ✗

Project students are due on - ✓ ✗

If we need to discuss, let me - ✓ ✗

FW: 'Publication Plans 09-10'. ✓ ✗

Please tell me your publicati - ✓ ✗

dmollaaliod@gmail.com (2)

Dominique Estival (1)

Duncombe, Deanne (CSIRO IMT (1)

Paris, Cecile (ICT Centre, Marsfi (1)

Tom Rowlands (1)

Figure 6.7: The Requests Pane showing grouping of tasks by person.

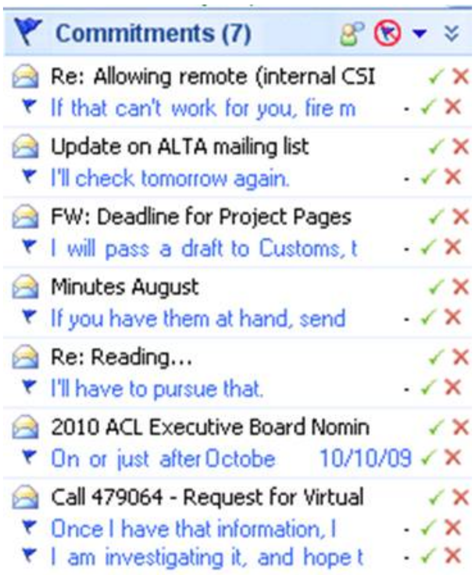
Requests can also be managed at either the message or sentence level; the user can click on the green tick icon next to an action, as shown in Figure 6.7, to mark that action as complete. Alternately, a user may delete a previously identified task by clicking the red cross icon.

The Requests Pane is populated with both incoming requests from other email senders, and outgoing commitments made by the mailbox user. Outgoing commitments are included in the request pane, since they place the same obligation on the mailbox user as an incoming request, as we discussed in detail back in Section 3.1.

By default, requests are displayed in reverse-chronological order. This ordering can be inverted, or the requests can be grouped by the person to whom the task relates, as in Figure 6.7. The person with whom a task is associated in this view is either the sender who made the request, or the recipient to whom the mailbox owner made the commitment.

6.1.2.3 Commitments Pane

The **Commitments** Pane shows the outstanding tasks waiting for someone else's action. As with the Requests Pane, these include both incoming commitments from other email senders, along with outgoing requests sent by the mailbox user which also create obligations that the mailbox user may wish to follow up on. The aim is to help the mailbox user keep track of actions that they are waiting on other people to fulfil.



Commitments (7)

Re: Allowing remote (internal CSI ✓ ✕
If that can't work for you, fire m - ✓ ✕

Update on ALTA mailing list ✓ ✕
I'll check tomorrow again. - ✓ ✕

FW: Deadline for Project Pages ✓ ✕
I will pass a draft to Customs, t - ✓ ✕

Minutes August ✓ ✕
If you have them at hand, send - ✓ ✕

Re: Reading... ✓ ✕
I'll have to pursue that. - ✓ ✕

2010 ACL Executive Board Nomin ✓ ✕
On or just after Octobe 10/10/09 ✓ ✕

Call 479064 - Request for Virtual ✓ ✕
Once I have that information, I - ✓ ✕
I am investigating it, and hope t - ✓ ✕

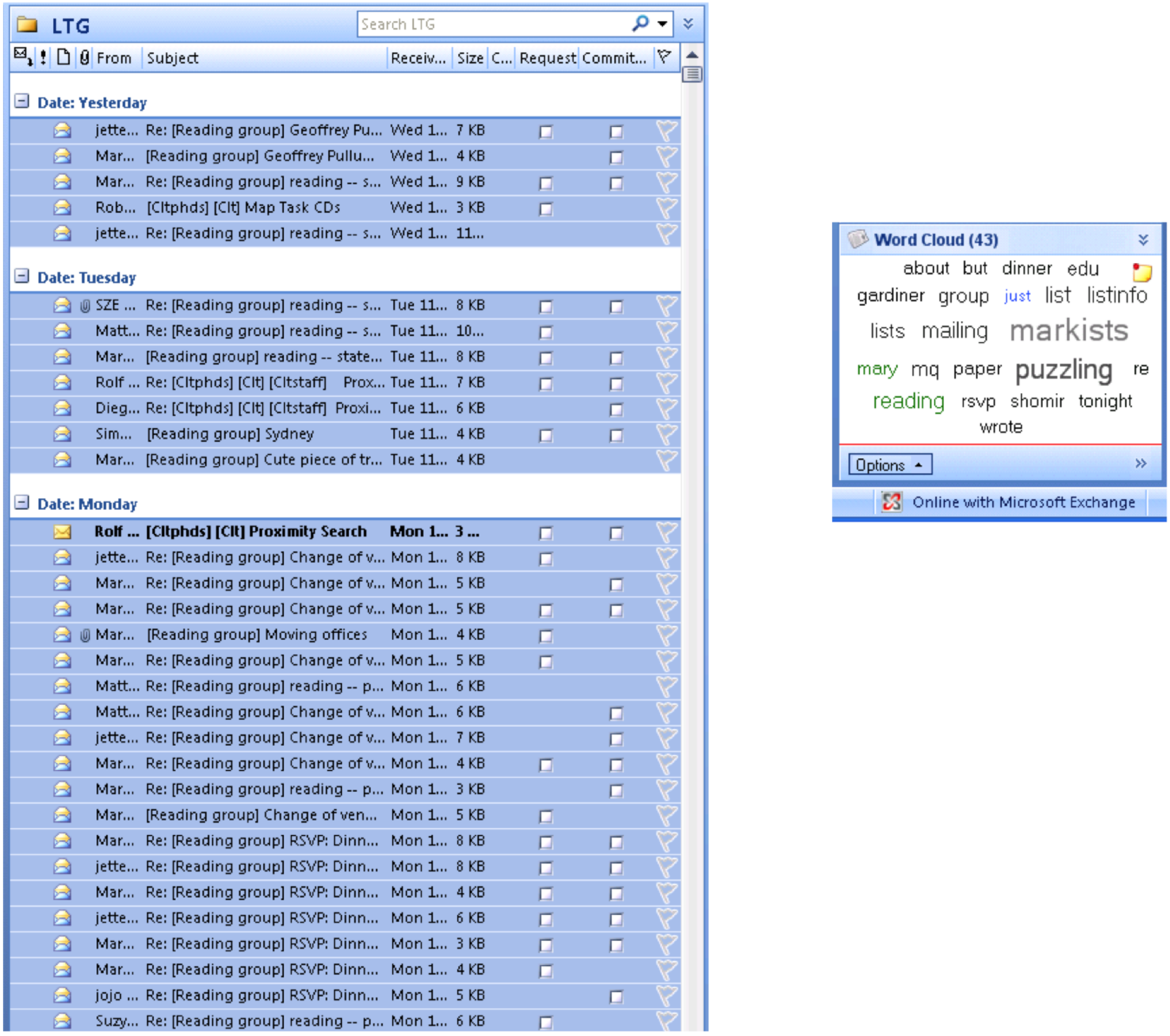
Figure 6.8: The Commitments Pane.

Figure 6.8 shows a chronologically ordered list of outstanding commitments extracted from the user's incoming and outgoing mail. As with the Requests Pane, commitments can be displayed at the message or utterance level, and can be optionally grouped by the person responsible for fulfilling the outstanding action.

6.1.2.4 Search and Word Cloud Panes

There is no existing capability in Microsoft Outlook, or any other email software that we are aware of, to search specifically within obligation content, separately from other email text. The **Search** pane in our plug-in contains a familiar search box that allows the user to search the content of all requests and commitments that have been either manually or automatically identified. Matching messages are presented back in a standard list, ranked in reverse-chronological order. This means that messages containing the most recent obligation act content that matches the search query are displayed first. The search functionality is intended to bring more of the features of a separate, structured task list to the email client, without removing the tasks from their original email context.

Separately, the **Word Cloud** pane, illustrated in Figure 6.9, shows a word cloud visualisation (Kaizer and Hodge, 2005; Halvey and Keane, 2007; Sinclair and Cardew-Hall, 2008) of words contained in the message(s) which are currently selected in the main message window in Outlook, which is shown in the left-hand pane of Figure 6.1. Word clouds are a popular way of visualising the frequency of words in a document or other textual content. We use the word cloud to show the frequency of words within both obligation and non-obligation content in the selected messages. Words that occur



LTG

Search LTG

From	Subject	Receiv...	Size	C...	Request	Commit...
Date: Yesterday						
jette...	Re: [Reading group] Geoffrey Pu...	Wed 1...	7 KB		□	□
Mar...	[Reading group] Geoffrey Pullu...	Wed 1...	4 KB			□
Mar...	Re: [Reading group] reading -- s...	Wed 1...	9 KB		□	□
Rob...	[Cltphds] [Clt] Map Task CDs	Wed 1...	3 KB		□	
jette...	Re: [Reading group] reading -- s...	Wed 1...	11...			
Date: Tuesday						
SZE ...	Re: [Reading group] reading -- s...	Tue 11...	8 KB		□	□
Matt...	Re: [Reading group] reading -- s...	Tue 11...	10...		□	
Mar...	[Reading group] reading -- state...	Tue 11...	8 KB		□	□
Rolf ...	Re: [Cltphds] [Clt] [Cltstaff] Prox...	Tue 11...	7 KB		□	□
Dieg...	Re: [Cltphds] [Clt] [Cltstaff] Proxi...	Tue 11...	6 KB			□
Sim...	[Reading group] Sydney	Tue 11...	4 KB		□	□
Mar...	[Reading group] Cute piece of tr...	Tue 11...	4 KB			
Date: Monday						
Rolf ...	**[Cltphds] [Clt] Proximity Search**	**Mon 1...**	**3 ...**		□	□
jette...	Re: [Reading group] Change of v...	Mon 1...	8 KB		□	
Mar...	Re: [Reading group] Change of v...	Mon 1...	5 KB			□
Mar...	Re: [Reading group] Change of v...	Mon 1...	5 KB		□	□
Mar...	[Reading group] Moving offices	Mon 1...	4 KB		□	
Mar...	Re: [Reading group] Change of v...	Mon 1...	5 KB		□	
Matt...	Re: [Reading group] reading -- p...	Mon 1...	6 KB			
Matt...	Re: [Reading group] Change of v...	Mon 1...	6 KB			□
jette...	Re: [Reading group] Change of v...	Mon 1...	7 KB			□
Mar...	Re: [Reading group] Change of v...	Mon 1...	4 KB		□	□
Mar...	Re: [Reading group] reading -- p...	Mon 1...	3 KB			□
Mar...	[Reading group] Change of ven...	Mon 1...	5 KB		□	
Mar...	Re: [Reading group] RSVP: Dinn...	Mon 1...	8 KB		□	□
jette...	Re: [Reading group] RSVP: Dinn...	Mon 1...	8 KB		□	□
Mar...	Re: [Reading group] RSVP: Dinn...	Mon 1...	4 KB		□	□
jette...	Re: [Reading group] RSVP: Dinn...	Mon 1...	6 KB		□	□
Mar...	Re: [Reading group] RSVP: Dinn...	Mon 1...	3 KB		□	□
Mar...	Re: [Reading group] RSVP: Dinn...	Mon 1...	4 KB		□	
jojo ...	Re: [Reading group] RSVP: Dinn...	Mon 1...	5 KB			□
Suzy...	Re: [Reading group] reading -- p...	Mon 1...	6 KB		□	

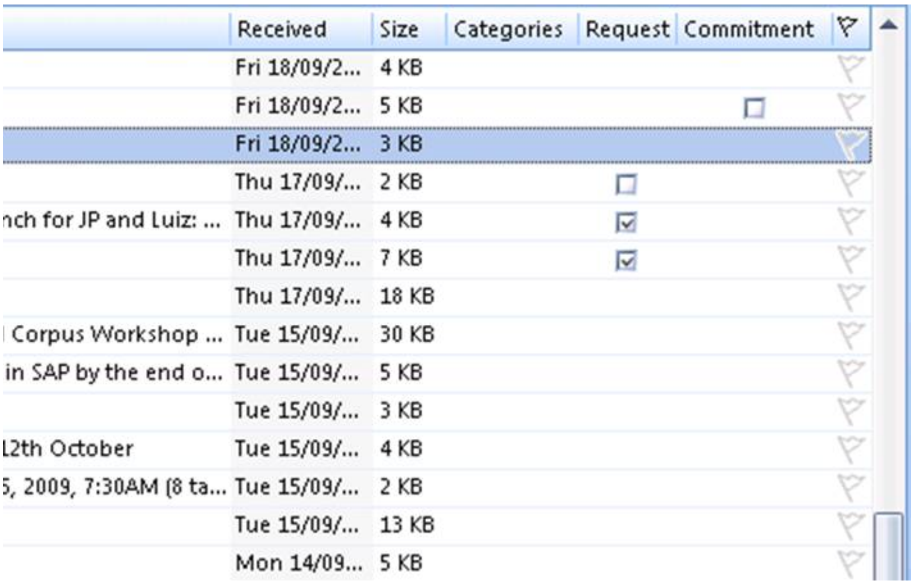
Word Cloud (43)

about but dinner edu gardiner group just list listinfo lists mailing markists mary mq paper puzzling re reading rsvp shomir tonight wrote

FIGURE 6.9: The Word Cloud for a group of selected messages. Frequent request words are in blue; frequent commitment words are in green; and frequent non-obligation words are in black.

frequently in requests are shown in green, words that occur frequently in commitments in blue, and other frequent words that occur in non-obligation act content are shown in black, as illustrated in Figure 6.9. In that example, we see *just* as a frequent commitment word, *mary* and *reading* as frequent request words, and remaining black words such as *markists* and *puzzling* as frequent words that occur in non-obligation acts within the selected messages on the left-hand side of Figure 6.9. We do some simple stemming and filtering of stop words to try to keep the display as relevant as possible.

Within the email plug-in, the word cloud is intended to function as an alternate summary of a single message or a collection of messages that provides a general overview of both their obligation content, and the context around the requests and commitments. Clicking on any word in the word cloud opens the message(s) in which the word occurs.



	Received	Size	Categories	Request	Commitment
	Fri 18/09/2...	4 KB			
	Fri 18/09/2...	5 KB			□
	Fri 18/09/2...	3 KB			
	Thu 17/09/...	2 KB		□	
nch for JP and Luiz: ...	Thu 17/09/...	4 KB		☑	
	Thu 17/09/...	7 KB		☑	
	Thu 17/09/...	18 KB			
Corpus Workshop ...	Tue 15/09/...	30 KB			
in SAP by the end o...	Tue 15/09/...	5 KB			
	Tue 15/09/...	3 KB			
l2th October	Tue 15/09/...	4 KB			
5, 2009, 7:30AM (8 ta...	Tue 15/09/...	2 KB			
	Tue 15/09/...	13 KB			
	Mon 14/09...	5 KB			

Figure 6.10: The Folder View showing request and commitment fields.

6.1.3 Folder View Indicators

The **Folder View Indicators** (Figure 6.10) integrate request and commitment information into the standard Outlook folder views. Message-level requests and commitments may also be annotated within the folder view using the toolbar buttons shown at the top of Figure 6.1.

When a user views their inbox or folder content, two additional columns are displayed. These indicate the presence of requests and commitments in each message. Because of the nature of the interface, only message-level presence is indicated; no detail of the specific request and commitment text spans is shown. The field indicators are three-state, indicating that a message contains either:

1. No Request/Commitment: a blank Request or Commitment field;
2. One or More Unfulfilled Requests/Commitments: an unchecked tickbox in the appropriate field; or
3. One or More Completed Requests/Commitments: a checked tickbox in the appropriate field.

These fields allow folders or other collections of messages to be easily sorted based on their request or commitment content and enable the user to quickly obtain an overview of the outstanding tasks in even a large collection of messages.

6.1.4 Summary

Together, the features of the Outlook plug-in are designed to allow users to fluidly and simply identify, manage and track unfulfilled requests and commitments across their incoming, outgoing and stored email messages.

We envisage that the novel tools provided for viewing, searching and navigating task-related content should lead to greater effectiveness in managing ongoing tasks, and fewer requests and commitments that remain unintentionally unfulfilled. We installed the plug-in for four individual email users, and their feedback was positive, especially in terms of the flexibility with which they could identify and view tasks in their email. None, however, used the plug-in for the extended period of time that would be required for more ecologically-valid evaluation. A formal evaluation of the effectiveness of the plug-in is left to future work.

6.2 Instrumenting and Logging User Actions

In order to understand how users interact with tasks in their email and to improve the performance of our automated request and commitment classifiers, we have instrumented many features of our plug-in to record user feedback and interaction.

6.2.1 Logging User Actions

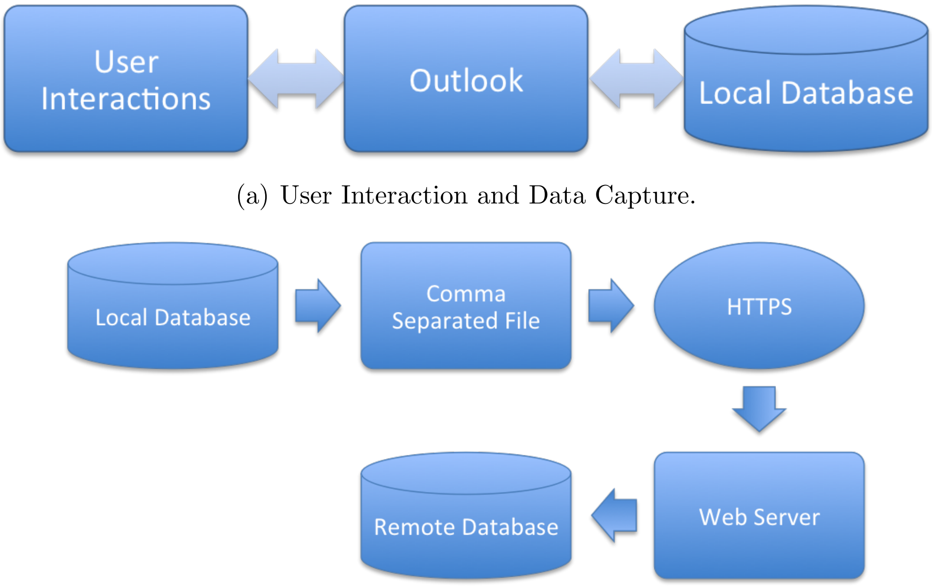
Most user interactions with our plug-in are instrumented and recorded. For example, each time a user identifies a request or commitment, this is recorded, along with metadata about the identified obligation and the surrounding message. We also log the specific interaction within the plug-in that resulted in the identification (e.g., whether the message-level button was pressed).

In terms of user actions, we log:

- all request and commitment annotations made by the user, either at the message level or at the text span level;
- all request and commitment deletion actions, which we interpret as corrections if a user deletes an automatic annotation;
- all request and commitment completion actions;
- message send and receive events;
- email message opening actions; and
- email reply actions.

6.2.2 Logging Process

Our logging framework for capturing this data is described in Figure 6.11. Figure 6.11(a) shows that we first capture user interactions from the Outlook plug-in and record this information into a local database, stored on the user's hard disk. Information is persisted in this way, because we cannot rely on users always having network



(a) User Interaction and Data Capture.

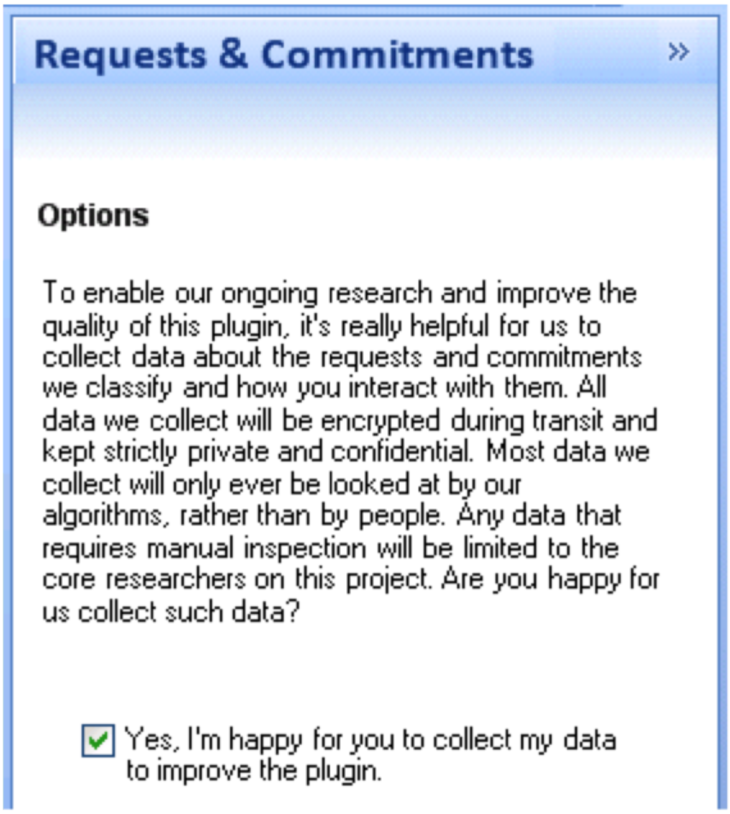
(b) Remote Network Data Upload.

Figure 6.11: The processes for data capture and data transfer that underpin the logging for the Outlook plug-in.

access while reading and interacting with their email. Outlook has wide-ranging support for offline email usage, with the ability to synchronise changes and send/receive new messages once connectivity is again available.

When the plug-in detects that network connectivity is available, Figure 6.11(b) shows how we periodically package up all the persisted user data since the last confirmed log delivery, and sends it as structured data (in CSV format) over an encrypted HTTPS session to our remote secure log server. At the log server, the received data is persisted into a central database that captures usage data for all email users who have the plug-in installed and logging enabled. This allows us to aggregate observations across a collection of users, as well as to drill-down into specific behaviour for an individual user.

Once we have persisted the user's interaction data to our central data store, we execute two additional steps to use this data to refine our classifiers: extracting annotations; and using this new data to retrain the request and commitment classifiers. This can be done on both a user-specific and a global level. Initially, we can tune the request and commitment classifiers to suit the data and preferences of the specific user. This process is not automatically triggered in the current plug-in, but could be improved to retrain at regular intervals (e.g., every night), or after receiving a threshold number of new pieces of implicit or explicit feedback on obligation acts. Over time, we can also build better baseline classifiers for all users based on the aggregated user data that is collected. We discuss these steps for retraining classifiers in the Section 6.3.1,



Options

To enable our ongoing research and improve the quality of this plugin, it's really helpful for us to collect data about the requests and commitments we classify and how you interact with them. All data we collect will be encrypted during transit and kept strictly private and confidential. Most data we collect will only ever be looked at by our algorithms, rather than by people. Any data that requires manual inspection will be limited to the core researchers on this project. Are you happy for us collect such data?

☑ Yes, I'm happy for you to collect my data to improve the plugin.

Figure 6.12: The Logging Opt-in Screen shown to the user after plug-in installation.

where we focus on use of the plug-in for email task management, and how we can improve performance through iterative feedback.

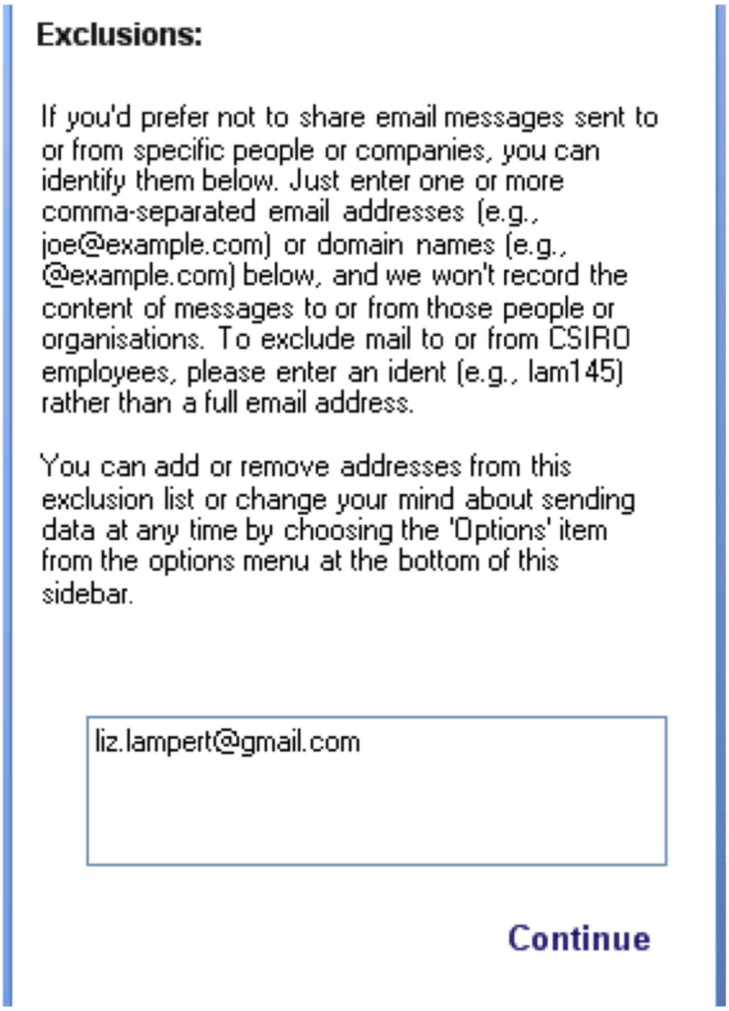
Importantly, while we have broadly instrumented the plug-in, we have implemented this logging carefully and deliberately to preserve the user's control and privacy. Below, we discuss the details of the privacy-preserving plug-in features.

6.2.3 Privacy and User Control

Privacy is an extremely important consideration when dealing with email data. Our desire to gather useful data from plug-in users must be balanced against the need to adequately protect people's privacy and to leave people in control of their data.

The plug-in continually logs interaction data into a local database on the user's computer. If the user opts-in to allow us to collect their usage data, we periodically transfer a batch of this data to a remote logging server for aggregated analysis.

To provide users with control over their data, we offer a range of opt-in and opt-out mechanisms. When the user first installs the plug-in, they have the choice to opt-in to allow their interactions and message content to be logged; the full opt-in text is shown



Exclusions:

If you'd prefer not to share email messages sent to or from specific people or companies, you can identify them below. Just enter one or more comma-separated email addresses (e.g., joe@example.com) or domain names (e.g., @example.com) below, and we won't record the content of messages to or from those people or organisations. To exclude mail to or from CSIRO employees, please enter an ident (e.g., lam145) rather than a full email address.

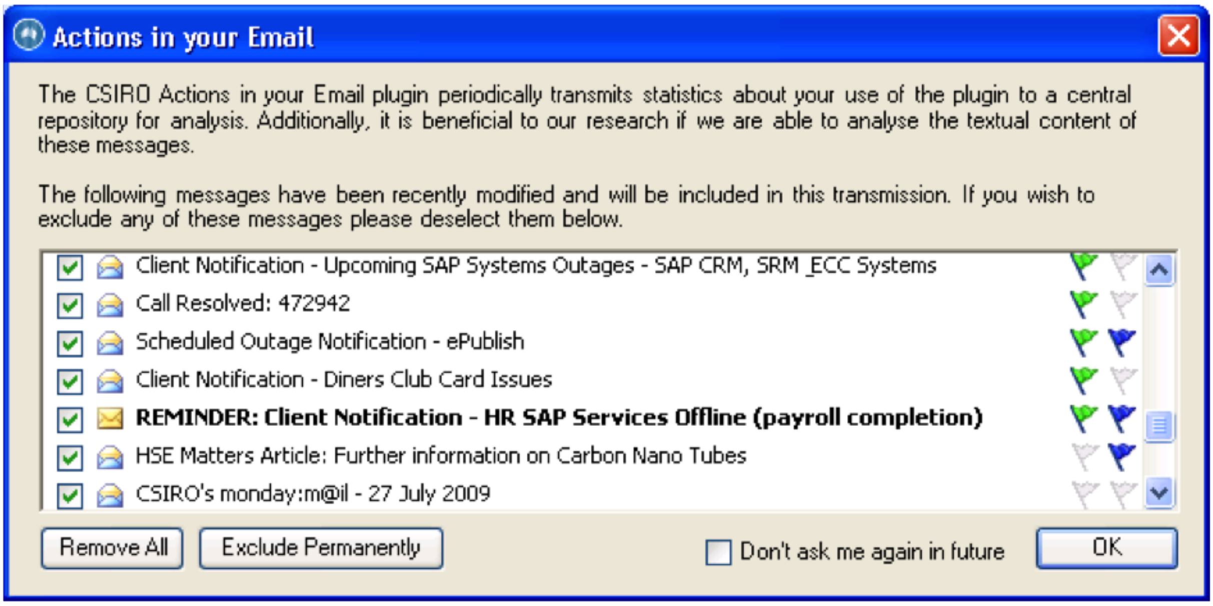
You can add or remove addresses from this exclusion list or change your mind about sending data at any time by choosing the 'Options' item from the options menu at the bottom of this sidebar.

liz.lampert@gmail.com

Figure 6.13: The Logging Exclusion Screen.

in Figure 6.12.

If the user chooses not to opt-in, no interaction data is remotely logged. If, however, the user opts-in to contribute their interaction and message data, we provide fine-grained opportunities to exclude specific messages. The first method is via an exclusion list, shown in Figure 6.13, that allows a user to exempt messages sent to or from particular people, email addresses or domains. All messages which match the exclusion list are not logged. The user is able to add or remove addresses from the exclusion list at any time. Additional privacy controls are provided in the Message View Panel; these allow the user to see whether the current message will be logged, and to choose to exclude the message from logging. If the message has already been remotely logged, then deselecting logging for the message, by clicking on the computer network icon shown at the right-hand side of the Message View Panel in Figure 6.2, will delete the logging data for that message at the next transfer of message data. The user can also



The CSIRO Actions in your Email plugin periodically transmits statistics about your use of the plugin to a central repository for analysis. Additionally, it is beneficial to our research if we are able to analyse the textual content of these messages.

The following messages have been recently modified and will be included in this transmission. If you wish to exclude any of these messages please deselect them below.

- ☑ Client Notification - Upcoming SAP Systems Outages - SAP CRM, SRM_ECC Systems
- ☑ Call Resolved: 472942
- ☑ Scheduled Outage Notification - ePublish
- ☑ Client Notification - Diners Club Card Issues
- ☑ **REMINDER: Client Notification - HR SAP Services Offline (payroll completion)**
- ☑ HSE Matters Article: Further information on Carbon Nano Tubes
- ☑ CSIRO's monday:m@il - 27 July 2009

FIGURE 6.14: The Remote Logging Confirmation Screen that details messages which are about to be remotely logged.

optionally add that sender's email address to their exclusion list to exclude all future email with that person from being logged.

Finally, prior to transferring any data to our remote server, the user is presented with a dialogue, shown in Figure 6.14, that confirms exactly which messages are about to be remotely logged. At this point, the user can inspect and exclude any or all of these messages before any data is transferred. Any data that the user agrees to contribute is encrypted during transit to the remote server, to ensure it cannot be easily intercepted. All logged data is stored securely, accessible only by the small group of researchers working directly on the Outlook plug-in project.

In future work, we plan to provide users with access to a secure web application through which they can login and view or delete any or all of their data that has previously been logged to our remote logging server.

6.3 Applications of the Outlook Plug-in

Having described the details of the plug-in features and instrumentation, in this section we discuss current and future use cases for which the plug-in has been developed: email task management; in-context email annotation; and extrinsic evaluation of novel email technology.

6.3.1 Using the Plug-in for Email Task Management

The primary drivers for integrating our request, commitment and zone classifiers within Microsoft Outlook is to allow them to be used for real-world email task management. The features described in Section 6.1 have each been specifically designed to provide task management support for email users. Our obligation act classifiers attempt to provide automated assistance in identifying tasks, and their predictions can be augmented by users manually identifying additional tasks, and correcting tasks which have been mistakenly identified.

Given the context-sensitivity and subjectivity around identifying requests and commitments,we install the plug-in with generically trained request and commitment classifiers for all users. Request and commitment classifiers are then retrained and adapted through ongoing feedback from each individual user to tailor the classification to suit their specific data and preferences.

One approach to gathering the feedback required for adapting our obligation act classifiers would be to employ detailed ethnographic interviews, coupled with physical observations and recordings of users in their natural work contexts. Unfortunately, this approach is impractical on any significant scale due to the resources required, and the limitations of not having physical access to users' real work environments for confidentiality, privacy and other reasons. Instead, we have focused on recording as much useful feedback as we can obtain automatically, based on the explicit and implicit user interactions with our plug-in, through our instrumentation and logging capability.

Unlike the RADAR system (Freed et al., 2008) that we discussed in Section 2.4.2.5, our plug-in has no model of the specific tasks or sequences of expected activity that occur in our users' email. This precludes our plug-in from offering the detailed, preemptive workflow support that RADAR offers. It does, however, allow our plug-in to function independent of any particular domains, making the plug-in applicable for any email user, regardless of the nature of their tasks. In this way, our plug-in shares a lot in common with the SmartMail system (Corston-Oliver et al., 2004) that we explored in Section 2.4.2.3. The SmartMail system presents automatically identified tasks to users, but does not allow users to create, delete or interact with tasks with the same richness that our plug-in provides. In this way, our combination of automated and manual task identification provides users with a level of flexibility beyond SmartMail and related systems.

In Section 6.2, we described the types of data that is captured, with the user's permission. This data and interaction logging functionality is an extremely important part of the plug-in. We can, for example, use data about automatically identified obligation acts that the user deletes as indicators of false positive classifications, and create additional training data that captures this insight. Similarly, we can use data about additional manual annotations to generate new training data that allows our classifiers to adapt the classifiers to this input. Task completions can be used as confirmation of correct automatic classifications, which can increase our confidence in those classification decisions. In all these cases, the logging and instrumentation within our plug-in allows user corrections and confirmations to be fed back to the classifiers to iteratively improve performance and to customise the classifiers to suit the users' specific preferences and data.

In this way, as well as providing a tool for improved task management in email, the plug-in also provides an in-context, manual annotation tool, which allows users to guide and improve the performance of our automatic request and commitment classifiers over time. We talk more about the use of the plug-in as a dedicated in-context annotation tool in Section 6.3.2.

Like our obligation act classifiers, our zone classifier is also deployed with a generically trained model. In contrast, however, we envisage less need to tune the zone classifier to specific users, and have no mechanism in the plug-in to capture user annotations of email zones, nor to iteratively refine and retrain our zone classifier. The zone classifier is, of course, deployed to help guide our request and commitment classifiers to consider only content from relevant email zones when automatically identifying obligation acts.

In future work we intend to employ active learning methods (Tong and Koller, 2002) to pro-actively prompt users to annotate specific messages that would provide maximum information to the automated request and commitment classifiers as training data.

6.3.2 Using the Plug-in for In-Context Email Annotation

We can extend the idea that we discussed above in Section 6.3.1 of using the plug-in for individual correction and tailoring of classifiers to think of the plug-in as a dedicated annotation tool. As we have discussed throughout this thesis, identifying requests and commitments (or other speech acts) is fundamentally concerned with pragmatic aspects of language. These are generally distinguished from syntactic and semantic aspects: syntax is concerned with the structure of language, semantics with its meaning, and pragmatics with language use within a given context (Morris, 1946). This context includes the linguistic context, the physical context in which the utterance is made, the personal, social and organisational context including the relationship between the email sender and recipient(s), along with other cultural and cognitive contexts, including the shared background knowledge of the participants in the interaction. All these elements figure into the appropriate pragmatic interpretation of an utterance; this dependence on context thus significantly increases the difficulty of classifying requests and commitments. Importantly, pragmatic annotation tasks such as the identification of requests and commitments, by the very nature of these tasks, can never be as reliably performed without access to such contextual information.

To gather human interpretations of the requests and commitments for this thesis, we have reported on a series of human annotation experiments. These experiments were conducted using the Enron email corpus, since it is the only large-scale collection of real-world email messages that is publicly available for research use. Together with annotation experiments performed by other researchers, these experiments have demonstrated that asking human annotators to identify particular speech acts (such as requests and commitments) in an isolated annotation task using a generic dataset such as the Enron email corpus can lead to considerable inter-annotator disagreement (Scerri, Handschuh, and Davis, 2008; Ulrich, Murray, and Carenini, 2008; Lampert, Dale, and Paris, 2008a).

A major reason for such disagreement is the out-of-context nature of the annotation task: the annotators do not have access to the wider context around the email messages in the Enron corpus as they were originally created and consumed. This makes interpreting and classifying the text of email messages much more difficult and ambiguous. In the absence of the real-world context, annotators bring their own differing assumptions and interpretation biases to the task, none of which may be obviously right or wrong. Thus, while using the Enron corpus provides real-world email data and allows other researchers to replicate our experiments, it also leaves annotators to guess at the context when interpreting and annotating messages. Additionally, the out-of-context nature of the Enron messages means we are unable to extrinsically evaluate the utility of the requests and commitments we identify, in terms of how useful the results are in helping people manage their real-world email tasks.

One solution to these problems is to capture the intuitions of annotators as they work with real emails in contexts where they are fully aware of their meaning and import. By integrating an annotation tool into Outlook, the plug-in provides an ecologically-valid environment that can be used for both out-of-context annotation experiments, such as those we have performed using the Enron email data, and in-context experiments that can use email messages from the annotators' own email mailboxes. We refer to this as in-context annotation, because the user understands the full context around the messages, including the people involved, the subject matter, the prior history of interaction, the environment in which the messages are being interpreted, and the relative importance of any tasks identified. When considering email messages in their own inbox, annotators are able to implicitly consider all of these contextual factors in assessing the intention and importance behind utterances in incoming email messages. This should lead to interpretations that are better able to reflect the real intentions of the sender and obligations for the recipients.

Working with annotators' personal email data, as in-context annotation implies, brings with it a host of challenges. Unlike in Section 6.3.1, where we looked primarily to exploit user data to improve and tailor classifier performance to suit the individual user, in-context annotation looks to gather and aggregate annotations from across users to look at issues of inter-annotator agreement and aggregated insights that can be drawn from in-context annotation of real-world email. This presents significant privacy challenges due to working with data that frequently contains confidential and personal information. There are also significant limitations in experimental replicability when working with such data, since each single annotator is likely to have access to different underlying email data, making it difficult to explicitly measure inter-annotator agreement, and to confirm or refute any experimental findings.

Importantly, however, in-context annotation of individual collections of email data provides the only 'natural' annotation environment that offers an ecologically-valid way of interpreting speech acts in email in a truly context-appropriate manner. More generally, the plug-in offers an example of the type of in-context tools that we believe are required for creating valid human judgements for complex document processing tasks.

6.3.3 Using the Plug-in for Extrinsic User Evaluation

Closely related to providing users with interfaces for annotating and interacting with requests and commitments in their email messages, the plug-in can also be used to observe and record how users interact with identified obligations within their email messages for the purposes of evaluation. Through measuring user actions and feedback, the plug-in provides a platform to both intrinsically and extrinsically evaluate automated request and commitment classifiers.

More importantly, we can start to gauge the actual effectiveness and utility of our classifiers, in ways that are not possible using the evaluation techniques we employed in Chapter 4 and 5. We can see, for example, how often automatically identified requests or commitments are explicitly or implicitly confirmed or corrected. Such confirmation can be explicit (the user re-identifies the same task), or implicit (the user marks the task as complete). Similarly, deletion of a task may indicate a mistakenly identified request or commitment. Measurements such as these start to tell us about the accuracy of our classification predictions in a real-world context, rather than measured against an ideal, but never perfect, gold-standard corpus.

The utility of the requests and commitments that we identify can also start to be assessed, through a combination of implicit and explicit user feedback. There are, however, challenges that remain in interpreting user actions as indicators of classification correctness and usefulness. If a user never interacts with a request that has been automatically identified, for example, what does that tell us about that classification prediction? Perhaps the obligation is spurious or unimportant, or perhaps the user has already dealt with the obligation. Without further information, it seems impossible to discern. What if a user adds a due date to a request and then later marks that request as complete? Does that confirm that the task was a useful one to identify? It seems likely that if a user goes to the effort of adding a due date to an obligation, that this is confirming the validity of that obligation. Similarly, marking a request as complete might validate that the task was relevant and now complete. There are other possible approaches to reduce such ambiguity. We could, for example, occasionally ask the user explicitly about whether a specific request or commitment was worth flagging.

There is no doubt that much more work is required before we develop sufficient understanding to be able to mitigate the inevitable ambiguity that arises from mapping user actions within the plug-in to observable feedback that we can use to evaluate the overall performance of our classifiers, or other novel email technology. It seems clear, however, that there is great potential to use the plug-in to assess the real-world utility and accuracy of our request and commitment classifiers in ways that are impossible without it.

In future work, we are also interested in exploring higher-fidelity evaluations that make use of the plug-in. We can side-step ambiguity in interpreting user actions by performing ethnographic observations of people as they work. This would provide an even richer context around evaluating how people work with tasks in email and the accuracy and utility of our automated request and commitment classifiers.

6.4 Summary

The plug-in we have developed for Microsoft Outlook provides an environment for deploying the range of request, commitment and zoning classifiers which we have developed in this thesis. We focused primarily on its use as an email task management environment to assist users in working with task content in email. We also recognise, however, that there are other use cases that this plug-in enables, including in-context annotation to capture the subtle and context-sensitive aspects of pragmatic meaning, and the extrinsic evaluation of the accuracy and utility of our tools for email task management, or indeed for other novel email tools.

In future work we aim to complete a user-centred evaluation of the effectiveness of our classifiers, and potentially of other techniques that we develop for presenting and interacting with obligation content in email. One specific such experiment will be to assess whether guiding the selection of text towards obligation content for inclusion in automatically generated summaries of email messages provides a more useful overview than the existing, first *n*-character email summaries that feature in Microsoft Outlook. Also left to future work are a range of additional technology advances, including an exploration of techniques for rephrasing extracted request and commitment sentences to provide better standalone task list entries, along with experimenting with approaches for the automatic extraction of task due dates.

"I dedicate somewhere between minimally two hours at the outlying range, up to ten hours on any given day trying to stay on top of email"

Email user, quoted in (Whittaker and Sidner, 1996, p. 277)

7
Conclusions, Limitations and Future Work

The identification of requests and commitments (or other speech acts) is fundamentally concerned with pragmatic aspects of language. As discussed in Chapter 2, these are generally distinguished from syntactic and semantic aspects: syntax is concerned with the structure of language, semantics with its literal meaning, and pragmatics with language use within a given context (Morris, 1946). Within pragmatics, it is widely recognised that it is very difficult to establish a definitive set of surface forms that correlate with a particular speech act (Searle, 1975). Winograd and Flores (1986, p. 159) went further, and posited that it is actually impossible to formulate a precise correspondence between the surface forms of utterances in a conversation and the structure of obligations that are conveyed. This is largely because, in the normal course of human communication, the complete intent of any piece of discourse cannot be determined by lexical or semantic analysis alone, but must be determined by context. This is especially true for indirect speech acts, which are particularly common in email text (Hassell and Christensen, 1996).

Throughout this thesis, we have worked to combat these challenges to establish how request and commitment speech acts, which form the basis of task-oriented communication, are realised and used in workplace email communication. Our aim throughout has been to build on this analysis of requests and commitments to create practical tools that assist email users to identify, manage and react or respond appropriately to obligation acts in both incoming and outgoing email messages.

Our approach has been built on detailed manual analysis of real-world email data, driven by a series of annotation experiments that we have used to bridge the gap of indirect mapping between surface forms and pragmatic meaning. These collections of human judgements about requests and commitments have also allowed us to analyse and account for the significant complexity that lies in how people actually exchange requests and commitments, with nuances of indirectness, context-sensitivity and ambiguity that we have analysed and worked to accommodate.

Building on the insights drawn from our annotated corpora, we have sought to design, build and evaluate effective tools and systems that can automatically identify requests and commitments within email messages. We have done this at the message, paragraph and sentence levels, assisted by our zone classifier which we use to understand the functional segments within each message before we seek to identify task-related content. Particularly at the finer levels of granularity, we had to deal with significant imbalance in our training and testing datasets: challenges which, to some extent, we have mitigated using techniques for controlling recall and precision, and through a novel system for building ensembles that combine classifiers at different granularities. As part of this work, we have demonstrated that ensembles of fine- and coarse-grained classifiers working together can outperform the same fine-grained classifiers working independently. This is an important result that has implications for other classification tasks which have different hierarchical levels at which classification can be performed.

Finally, we have applied and integrated our message, paragraph, sentence and zone classifiers into Microsoft Outlook, one of the most widely used software products for business email. Our goal with this work has been three-fold: to increase the visibility and role of tasks within email clients relative to commonly used commercial email applications; to provide an environment for in-context annotation; and to provide a platform for future in-context, extrinsic evaluation of novel email technology.

Together, this thesis offers a comprehensive account of techniques for identifying requests and commitments in workplace email communication.

7.1 Summary of Thesis Contributions

As discussed above, this thesis forms a body of work that explores the nature of requests and commitments in email. Together with this analysis, we have also designed, built and evaluated technology for automating their detection in real-world, workplace email. Within this body of work, there are a range of specific contributions which we hope will prove useful for other researchers in ongoing email research. We discuss the major contributions below.

Taken together, our contributions of detailed analysis, annotated corpora, automated classifiers and prototype integration software provide a well-grounded foundation for other researchers to conduct further empirical task-related email research with real-world data and in real-world contexts.

7.1.1 Empirical Definitions of Requests and Commitments

Throughout Chapter 3, we presented a detailed analysis of real-world usage of requests and commitments in workplace email. Our insights there are drawn from human judgements across a range of annotation experiments. As detailed in Chapter 3, we consider, for example, that offers and invitations can function as commitments, but threats usually do not. We also consider that both one-time and ongoing requests and commitments should be identified, as should requests and commitments for inaction. Phatic obligation acts, however, should either be ignored or marked as distinctly different phenomena from actionable requests and commitments. The resulting set of definitions account for the complexity observed in actual usage of requests and commitments in a more comprehensive manner than has been done in related work in the literature.

7.1.2 Automated Request and Commitment Classifiers

In Chapters 4 and 5, we presented a series of classifiers that automatically identify requests and commitments at the message, paragraph and sentence levels.

Our message-level classifiers can identify requests with an accuracy of 83.76% for requests and 86.62% for commitments. At the paragraph level, we achieved a peak weighted F-score of more than 0.91 for requests, and 0.80 for commitments; at the sentence level, we saw weighted F-scores of 0.88 for requests and 0.71 for commitments.

While it is always difficult to compare performance against systems that use different data for training and testing, our classifiers significantly outperformed the baseline systems, and reached F-scores that compare favourably with related work.

7.1.3 A Coarse-to-Fine-Grained Ensemble Approach to Classification

Part of the work we presented in Chapter 5 included a novel coarse-to-fine classifier ensemble to counter high levels of data skew. Our ensemble approach, which drew on our message, paragraph and sentence classifiers to improve performance, reliably increased the weighted F-score for automated commitment identification by 5–12% at the paragraph and sentence levels. Beyond the encouraging performance improvement we saw for commitment identification, our hope is that this technique may have application in related problems where there is information that can be better exploited through combining classifiers at different levels of granularity.

7.1.4 A System for Email Zoning

In Chapter 4, we developed and presented a novel system for automatically identifying functional zones within email messages. This system demonstrated cross-validation accuracy of more than 90% in two-zone and three-zone configurations, and accuracy of 87% when classifying all nine zone types that we proposed. Even more importantly, our evaluation results provided evidence that performing automated zoning prior to identifying requests or commitments in an email message can improve request and commitment classification accuracy by 5–16%.

7.1.5 A 1000 Message Annotated Email Zoning Corpus

As the basis of our email zoning work, we developed an annotated corpus of 1000 email messages, in which each line of text was annotated with one of nine zone categories. This annotated email zone corpus has been made publicly available for research use, under a Creative Commons licence,[1] and as of July 2013 has been downloaded by almost 100 different people and organisations since its release.

7.1.6 A 1000 Message Span-Annotated Email Obligation Act Corpus

To support our automated request and commitment classification work, as well as to provide the corpus data for our manual analysis of these phenomena, we conducted a series of annotation experiments. The annotated email corpora from these experiments provided us with a rich collection of pragmatically annotated data, culminating in a corpus of 1000 email messages that was annotated by two annotators at the text span level. This span-annotated corpus has also been made publicly available for research use, under a Creative Commons licence.[2]

7.1.7 A Prototype Software Plug-in for Microsoft Outlook

To embed our research back in the real world of email users, we developed a prototype software plug-in for Microsoft Outlook. This plug-in integrated our message-level and sentence-level request and commitment classifiers into the world's most widely used business email software.

The plug-in provides users with an integrated task management capability that allows them to directly interact with and manage tasks, without having to leave their message context. The plug-in also offers a platform for evaluating the accuracy and effectiveness of automated classifiers in real-world contexts, and will enable extrinsic evaluation of the identification of pragmatic phenomena such as requests and commitments. As well as providing an environment for real-world use and evaluation of our classifiers, we have also designed the plug-in as a platform for in-context annotation of future email corpora, to provide higher fidelity and more context-aware judgements of email data for further improvement of our request and commitment classifiers. The plug-in is adaptable, in the sense that it can be used to deploy and evaluate other novel email technology.

7.2 Future Work

There are many avenues along which to continue and extend the work that we have presented in this thesis.

While we have extensively evaluated the intrinsic performance of our automated classifiers, we have not performed an extrinsic user evaluation. Such experiments could be conducted by deploying our Microsoft Outlook plug-in to a group of evaluation users, and observing behaviour and feedback over a series of controlled and open tasks. This evaluation could be completed with any novel email technology that can be integrated into the Microsoft Outlook plug-in.

[1] Available from: http://zebra.thoughtlets.org/data.php.

[2] Available from: http://zebra.thoughtlets.org/spans/data.php.

As noted in Chapter 2, unsupervised and semi-supervised methods of obligation act recognition are an area of increasing research. While current methods have significant limitations relative to supervised methods, there is a lot of scope to exploit implicit data from email users to develop better unsupervised and semi-supervised methods for recognising pragmatic phenomena such as patterns of obligation acts in email. Given the high expense, in both time and effort, of gathering manually labelled corpora of human judgements as we have done in this thesis, it is likely that these techniques will have an increasing role to play in future research.

There are also many related tasks which we have not focused on in this thesis. Such tasks include automated task deadline identification and extraction, determining the urgency and importance of identified tasks, and rewriting requests and commitments to be easily interpreted when extracted to a standalone task list. All these tasks provide further research challenges that could be readily explored by building on the work we present here.

Finally, there is an ongoing need to perform ethnographic studies of workplace email usage, to understand how people are adapting their behaviour in the face of changing technology and work practices. We have seen Fisher et al. (2006) revisit the seminal work of Whittaker and Sidner (1996) as an example of this type of work. There is clear scope for further study of work practices and task-oriented email usage inside and outside the workplace. Such studies provide the foundation on which work such as our own is based; ensuring that our understanding of email practices remains current will help ensure that future email research will remain relevant to the changing real world needs of email users.

A
Annotation Guidelines

This appendix contains a copy of the annotation guidelines that we provided to our human annotators. The guidelines below represent the instructions that were followed in generating the data used for training our computational algorithms for request and commitment detection in Chapter 4 and 5.

A.1 Span-Annotation Guidelines

Thank you for participating in this experiment.

You will be presented with a series of unconnected email messages, and should annotate every span[1] in each message which contains:

1. A request;
2. A commitment;
3. A phatic request;
4. A phatic commitment; or
5. Some combination of the above.

You may also record the reasons for your decisions, or other matters you consider worthy of note, in the Comments field. Definitions of requests, commitments and phatic obligation acts are provided below.

[1] The guidelines included are for our span-annotation task. Our annotation guidelines for message-level annotation differ in minor ways. These variations, and copies of earlier versions of annotation guidelines, are included in the discussion of our annotation experiments in Appendix B.

Please:

- Annotate on the basis of the text written by the current email sender, and not on the basis of included or replied-to text: in other words, ignore any requests, commitments, phatic requests or phatic commitments in text that appears in the content of earlier messages. Only consider such quoted content in order to interpret or clarify the intent of new content in an email message.
- Annotate from the point of view of the recipients for each email message: in other words, ask yourself whether at least one recipient would consider the message to contain a request or a commitment, or a phatic request or phatic commitment.
- Note that a text span may contain any combination of requests, commitments, phatic requests and phatic commitments.
- Provide explanation or justification for your annotation choice in the Comments field if you believe determining the presence of a request, commitment is particularly context-dependent or open to subjective interpretation.
- You may stop annotating and logout at any time. All submitted annotations will be saved. You can return and continue annotating at any time. Once you have read and understood the annotation instructions below, please begin.

A one-page summary of these guidelines[2] is also provided for you to print out and reference during the annotation task.

Requests

We explain what a request is below. There are various different kinds of requests, so the comments here are intended to demonstrate the variety of these; but note that you are not required to explicitly distinguish between these in any given instance. You only need to indicate whether a request of any of the described types is present in the email message.

1. A Request is an utterance from the email sender that places an obligation on the email recipient to:
   - schedule an action (e.g., by adding something to a calendar or task list);
   - perform (or not perform) an action; or
   - respond with some speech act (explained below).

[2] We include this one page guideline below in Section A.2 of this Appendix.

Examples:

(A.1) *Please come along to the meeting next Friday at 2pm.*

(A.2) *Make sure you finish the report today.*

(A.3) *Are you in the office tomorrow?*

2. A request to perform a speech act response may involve communicating with the email sender or another agent. Speech act responses can include any type of speech act, but often involve informing, permitting, committing, accepting and refusing.

Examples:

(A.4) *What time are we meeting today?*

(A.5) *Please let John know whether you'll be attending.*

(A.6) *Is it ok if I come in late on Friday?*

3. Requests may be conditional or unconditional.

   Unconditional requests require action or response without consideration of any preconditions.

Examples:

(A.7) *Please come along to the meeting next Friday at 2pm.*

(A.8) *Can you please send me the plan?*

4. Conditional requests only oblige action or response from the email recipient if some stated condition is satisfied. The stated condition is often in the same utterance as the request, but may be elsewhere in the email message.

Examples:

(A.9) *If you want comments, I need the project plan completed today.*

(A.10) *You'll need to finish the plan today if you want my feedback.*

5. Requests may also be direct or indirect. Note that this distinction is independent of the conditional-unconditional distinction.

   Direct requests state the request for action or response explicitly, either as an imperative command or as a question whose literal interpretation would result in a speech act that responds with the required content.

Examples:

(A.11) *Please send me the project plan today.*

(A.12) *What time is the meeting today?*

6. Indirect requests do not explicitly instruct the recipient to act, or do not explicitly instruct the recipient to respond with the required speech act; but an action is still expected.

Examples:

(A.13) *I need you to send me the project plan today.*

(A.14) *Do you know what time the meeting is today?*

(A.15) *I was wondering if you might be able to introduce me to John.*

7. Requests for inaction prohibit action or request a negated action, but are still considered to be requests.

Examples:

(A.16) *Please, don't distribute this: It's an internal document.*

8. Meeting Announcements are considered to be (indirect) requests, even if no explicit request to attend the meeting in question is made.

Examples:

(A.17) *Today's Prebid Meeting will take place in EB32c2 at 3pm.*

9. Messages with Attachments are considered requests to open, read, or act on the attached document(s) if, on the basis of the content of the email message, you believe you would feel obliged to open, read or otherwise act upon the attached file(s).

Examples:

(A.18) *Document for review is attached.*

(A.19) *Please open the attached document to view a list of products.*

(A.20) *Please review the attached document.*

10. Requests that commit a third party are only considered to be requests if the agent being committed is a recipient of the message. Mail aliases, mailing lists and group email addresses do not identify an individual as a recipient.

Examples:

(A.21) *Robert will send around an agenda before the meeting.*

The above utterance is a request if and only if Robert is a direct recipient of the email message. It is not a request if the message is sent only to a mail alias or mailing list, since it is not possible to determine the recipient set. Note that in either case, the above utterance is considered a commitment.

Commitments

We explain what a commitment is below. There are various different kinds of commitments, so the comments here are intended to demonstrate the variety of these; but note that you are not required to explicitly distinguish between these in any given instance. You only need to indicate whether a commitment of any of the described types is present in the email message.

1. A commitment is an offer or promise by the sender of the email for future action or response on behalf of an identified agent. The person on whom the commitment is placed is usually the sender, but it may alternatively be some other assigned person, group of people, or organisation. You might expect to find the action on the responsible person's task list or calendar.

   Examples:

   (A.22) *I'll be at the meeting next Friday at 2pm.*

   (A.23) *I can book a meeting room if you like.*

   (A.24) *Liz will book the meeting room for us.*

   (A.25) *Please let me know where you are located, and I'll try to get it over there today.*

   (A.26) *Our finance department will pay the invoice by the end of the month.*

2. If the agent committed to the future action is not identifiable, the utterance is not considered to be a commitment.

   Examples:

   (A.27) *The briefing will be in Room F on the first floor.*

3. Commitments may be conditional or unconditional.

   Unconditional commitments commit to future action without imposing any preconditions.

Examples:

(A.28) *Either Sean or I will forward the e-mails to you.*

(A.29) *I'll keep you posted on any changes.*

4. Conditional commitments only commit to future action if some stated condition is satisfied. They can often be considered as offers for future action. The stated condition is often in the same utterance as the commitment, but may be elsewhere in the email message.

Examples:

(A.30) *If you wish, I could provide you questions in advance.*

(A.31) *Katherine will give you the report if I'm not at my desk.*

5. Commitments may also be direct or indirect. Note that this distinction is independent of the conditional-unconditional distinction.

   Direct commitments explicitly state the commitment to future action.

Examples:

(A.32) *I'll book a room.*

(A.33) *I'll do it.*

(A.34) *John will book a room if we need one.*

6. Indirect commitments do not explicitly state the commitment to future action, but imply that it will take place. They may or may not state the future action.

Examples:

(A.35) *More details to follow.*

(A.36) *John can book a room.*

(A.37) *Leave the room arrangements to me.*

(A.38) *Leave it to me.*

7. Commitments to inaction are considered to be commitments.

Examples:

(A.39) *I won't send any further updates.*

(A.40) *John won't be in the office on Friday.*

Phatic Requests and Phatic Commitments[3]

1. Phatic requests are utterances that could function as a request in some other context, but are not requests in the context of use. Similarly phatic commitments are utterances that might function as a commitment in a different context of us. Phatic requests and phatic commitments are frequently formulaic in their form, and they do not place any significant obligation on the recipient or other identified agent to act or respond.

   Phatic requests are not requests and phatic commitments are not commitments.

   Examples:

   (A.41) *How are you?*

   (A.42) *Let me know if you have any questions.*

   Variations on 'Let me know if you have any questions' are particularly common in email messages. The context of the entire email message should be considered to distinguish between whether such utterances function as a request or commitment, or represent a phatic request or phatic commitment.

A.2 Single-Page Annotation Instructions

Below is an additional single-page guide that we made available to annotators as a convenient reference during annotation. It was designed to fit on a single A4 page to be printed and referred to during annotation.

Summary of Email Span Annotation Guidelines

1. Annotate each text span that expresses a request.
2. Annotate each text span that expresses a commitment.
3. Annotate each text span that expresses a phatic request.
4. Annotate each text span that expresses a phatic commitment.
5. Record reasoning and other observations for each annotation in the Comments field.

Annotate based on the content written by the current email sender, unless you feel strongly that obligation is implied from other message parts. By default, use quoted content only to interpret or clarify the intent of new content in an email message.

Annotate from the point of view of the set of identified recipients for each email message; mark all requests that place obligation on at least one recipient.

[3] In earlier versions of our annotation guidelines for message annotation we referred to phatic requests and commitments as PLEASANTRIES. The definitions were equivalent.

Requests

Any utterance from the sender that places an obligation on one or more recipients to schedule an action (e.g., by adding something to a calendar or task list), execute an action, or respond with some speech act is a request.

1. Can be Conditional or Unconditional.
2. Can be Direct and Indirect.
3. Include Requests for inaction.
4. Include Meeting announcements and Meeting requests.
5. Include Process instructions / Hypothetical Requests.
6. Include Attachment Requests

Do NOT consider as requests:

1. Phatic Requests.

Commitments

Any utterance from the sender that offers or promises future action or response for an identified agent is a commitment. The person on whom the obligation for future action or response is placed is often, though not always, the sender.

1. Can be Conditional and Unconditional.
2. Can be Direct and Indirect.
3. Include Commitments to inaction.

Do NOT consider as commitments:

1. Utterances where the agent committed to future action is not identifiable.
2. Statements about Future State without assigned responsibility.
3. Statements of intention or desire without promise or offer of future action.
4. Phatic Commitments

Phatic Requests and Commitments

Resemble the form of requests and/or commitments, but are often formulaic, their function is not to promise or elicit action or response.

Phatic requests and commitments are not considered to be requests or commitments. Please annotate them with the separate phatic annotation categories.

B
Annotation Experiments

This appendix contains further information about the series of annotation experiments that we conducted to gather insights into requests and commitments, and to generate gold-standard training data to train supervised machine learning algorithms.

A high-level overview of this series of experiments is described in Section 3.2.2. Here, we describe each annotation experiment in more detail, focusing on variations in the annotation guidelines provided to annotators in each experiment. The final set of guidelines that was used for span-level annotation is included in Appendix A.

B.1 Sentence 1: A Pilot Sentence-Level Annotation Experiment

The first annotation experiment was undertaken as a pilot annotation task, to develop and test annotations processes and guidelines. The data we used in this experiment consisted of 54 email messages containing 310 sentences that were manually extracted from the Enron email corpus. For this exploratory stage of annotation, selected messages were constrained to contain less than twelve sentences, to avoid the effort associated with coding longer messages. Email messages were selected to represent a variety of syntactic styles of expressing possible requests and commitments.[1] In particular, we attempted to select examples of sentences representing both explicit and implicit requests and commitments, and sentences with and without explicit task addressivity (i.e., sentences that do and do not address requests to a specific, named recipient).

[1] Note that in this early experiment, we referred to requests as REQUESTS-FOR-ACTION and commitments as COMMITMENTS-FOR-ACTION. We revised this terminology in later experiments, recognising that REQUESTS and COMMITMENTS are frequently more varied than this—e.g, requests-for-response, requests-for-permission, commitments-to-permit.

B.1.1 Processing Email Body Text

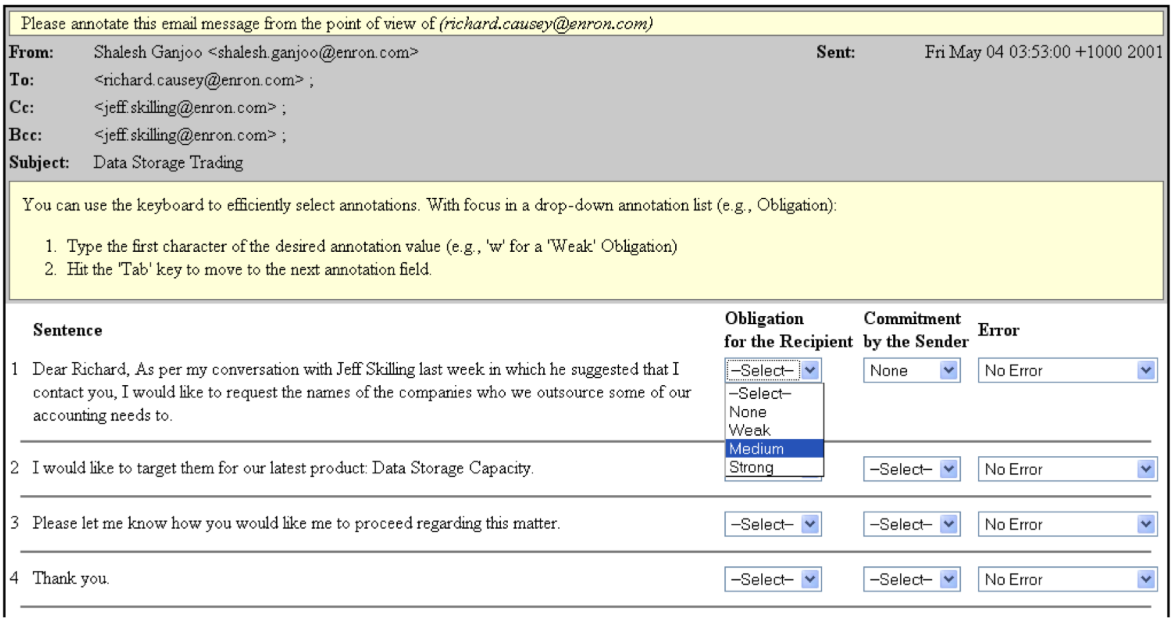
We automatically pre-processed the text in the body of each email message to try to remove email signatures and all quoted or forwarded email content, leaving only text written by the author of the current email message, also called AUTHOR TEXT. As we discuss in much more detail in Section 4.4, we refer to this task of segmenting email messages into different functional blocks of text as EMAIL ZONING.

At the time of running this first annotation experiment, we had not developed our Zebra system for email zoning. Instead we used the Jangada software (Carvalho and Cohen, 2004) to identify signature blocks and reply lines. In using this software, we identified similar shortcomings to those identified recently by Estival et al. (2007) who used Jangada on their own email corpus. In particular, Jangada did not accurately identify forwarded or reply material in the email messages we used for our annotation task. We posit that at least one factor in the poor performance of Jangada is due to it being trained on data from Usenet newsgroups, which generally uses different syntactic markers for forwarded and quoted material. We believe this is a significant factor in the systematic errors that Jangada makes in failing to identify quoted reply and forwarded content represented in the style used by Microsoft Outlook. Outlook is the most common email client used to compose messages in the Enron corpus. Given a lack of other available email processing tools, we used the Jangada software despite its shortcomings, and allowed annotators to flag processing errors during the annotation process. As described in Section 4.4, we subsequently implemented our own system to perform email zoning, and thus only used Jangada for this initial experiment.

B.1.2 Sentence Splitting

Because we wanted annotations at the sentence level, once we had attempted to remove quoted reply content, forwarded content and email signatures, we then segmented the body of each email message into sentences. For this purpose, we used the SentParBreak sentence and paragraph segmenter (Piao, Wilson, and McEnery, 2002). SentParBreak uses heuristic rules for identifying the boundaries of sentences and paragraphs. For this pilot task, we did not attempt to refine these rules for email data, and instead used SentParBreak without modification. In later experiments, we added further rules and heuristics to improve performance on email data. We applied SentParBreak to the bodies of all 250,000 email messages in our corpus and produced just over 3 million sentences of probable author text. Due to the Jangada processing errors, however, we know that some proportion of these sentences actually contain quoted reply or forwarded email content.

We did not formally evaluate the performance of the SentParBreak as a sentence segmenter; instead, we allowed annotators to flag sentence segmentation errors during the annotation process, using the same human-centred approach as applied for dealing with Jangada processing errors.



Please annotate this email message from the point of view of *(richard.causey@enron.com)*

From: Shalesh Ganjoo <shalesh.ganjoo@enron.com> **Sent:** Fri May 04 03:53:00 +1000 2001
To: <richard.causey@enron.com> ;
Cc: <jeff.skilling@enron.com> ;
Bcc: <jeff.skilling@enron.com> ;
Subject: Data Storage Trading

You can use the keyboard to efficiently select annotations. With focus in a drop-down annotation list (e.g., Obligation):

1. Type the first character of the desired annotation value (e.g., 'w' for a 'Weak' Obligation)
2. Hit the 'Tab' key to move to the next annotation field.

	Sentence	Obligation for the Recipient	Commitment by the Sender	Error
1	Dear Richard, As per my conversation with Jeff Skilling last week in which he suggested that I contact you, I would like to request the names of the companies who we outsource some of our accounting needs to.	–Select–	None	No Error
2	I would like to target them for our latest product: Data Storage Capacity.		–Select–	No Error
3	Please let me know how you would like me to proceed regarding this matter.	–Select–	–Select–	No Error
4	Thank you.	–Select–	–Select–	No Error

Figure B.1: The web-based annotation tool used in the *Sentence 1* annotation task.

B.1.3 Annotation Task Details

A primary goal of this pilot annotation task was to explore the level of human agreement in identifying requests and commitments within email messages. We gave three annotators a set of annotation guidelines, described in Section B.1.4, and asked them to independently identify requests and commitments within our corpus of 310 sentences from 54 email messages. The guidelines vary from our final guidelines in Appendix A, mostly in terms of containing less overall detail, attempting to define different strengths of request and commitment, and lacking guidance around specific phenomena such as how to interpret conditional requests and commitments.

The pilot annotation task was performed using a custom web-based tool, developed using Ruby-on-Rails, shown in Figure B.1. This annotation tool displays email messages using a look-and-feel that approximates the way email messages are displayed in Microsoft Outlook. Each email message is presented with the usual header fields and values: From, Date, To, Cc, Bcc, and Subject. Below the header information, in the email content pane, the author text of the email message was presented as a sequence of sentences to be coded. Paragraph breaks in the original email were represented by a single uncodable blank line in the annotation tool.

For each sentence, annotators were required to select annotation values from a number of aligned drop-down menus. Using these menus, annotators performed three actions for each sentence:

1. First annotators indicated whether the sentence expressed a request for the specified recipient, and, if so, whether the request was weak, medium or strong. (See

below for an explanation of how we defined the specified recipient.)

2. Next, annotators indicated whether the sentence expressed a commitment from the sender, and if so, whether the commitment was weak, medium or strong.

3. Finally, annotators could optionally flag any processing errors with the sentence. Flaggable processing errors included sentence segmentation problems, and the inclusion of quoted or forwarded email material (non author text).

Annotators were instructed to interpret each sentence in the context of the entire email message, rather than in isolation. At the top of each message, one recipient to whom the email message was originally sent – the SPECIFIED RECIPIENT – was noted, and annotators were instructed to approach the annotation task from the point of view of that person. This instruction is shown at the top of Figure B.1. For the purposes of this pilot annotation task, the first non-sender recipient of the message was chosen as the specified recipient. The 'non-sender' constraint was introduced to counter cases where the first recipient was actually the sender (presumably copying their own email to themselves for action or recall purposes). Where no explicit recipients were identifiable, as in the case of an email message whose recipients are all blind carbon copied, annotators were instructed to interpret the email message from the point of view of a generic recipient of the message. In general, requests and commitments in such messages tend to be addressed to all recipients, meaning that annotating from the point of view of a generic recipient is acceptable. Annotators were instructed not to mark any requests directed explicitly to recipients other than the specified recipient as requests. Our explicit instruction around whose point-of-view to adopt during annotation was intended to increase annotator agreement. Recognising that our task involved pragmatic-level interpretation and annotation, it was important to fix as much of the context as possible to encourage more consistent interpretation. Figure B.1 also shows some additional annotator instructions that were provided to guide new annotators during the first few messages that they attempted to annotate.

To qualify, explain or otherwise comment on any aspect of their interpretation, annotators were able to make notes using a comments field for each email message. This comments field was also used in combination with the OTHER category of processing errors to highlight processing or display problems other than segmentation and author text related issues. Annotators were instructed to use the comments field to explain any annotation decisions which they felt were conditional, subjective or particularly context-sensitive. For example, if an annotator's decision depended on potentially ambiguous interpretation of the email message or its context, they were instructed to explain the basis for their annotation.

Finally, we also noted to annotators that it was possible for a single sentence to contain both a request and a commitment (e.g., *Please send the document today, so I can get comments back to you by Monday*). The annotation tool made it possible for any sentence to be annotated as both a request and a commitment.

B.1.4 Annotation Guidelines

Annotators were instructed with the following definitions and guidelines for the pilot annotation task.

A request places some form of obligation on the recipient to respond or act. A simple test for a request is: Is this sentence asking me to do something? Examples of actions can include, but are not restricted to:

- answering a question, in email or otherwise;
- forwarding the message to a new recipient; or
- performing some action in the real world, such as preparing a document or gathering some data.

You should annotate every sentence that carries an expectation that the specified recipient of the email message should respond or take some action as a request-for-action. When you mark a request-for-action, you should also indicate its strength, as follows:

- **Strong:** Action or response from the specified recipient is considered important and/or mandatory.
- **Medium:** The sender expects a response or action from the specified recipient.
- **Weak:** Action or response from the specified recipient is optional or conditional; the sender would find it reasonable if the specified recipient took no action.
- **None:** No request is expressed.

Commitments by the sender represent a promise from the author that some future action will be taken. A simple test for a commitment is: Is this sentence promising to do something?

Commitments occur both when an action is to be taken by the sender, or when the sender promises action on behalf of another person. The reason for including such Commitments is based on the intuition that delegated promises might occur frequently because of the hierarchical nature of many workplaces, and are likely to be an important part of workplace conversations for action. An example is:

(B.1) *Peter will call to let you know the final arrangements.*

You should annotate any sentence that carries an expectation that the sender of the email message will take responsibility for some action being taken as a commitment.

As for requests, you should indicate the strength of each commitment, as follows:

- **Strong:** Action from the Sender is considered important and/or mandatory.
- **Medium:** The specified recipient expects a response or action from the Sender.
- **Weak:** Action or response from the Sender is optional or conditional; the specified recipient would find it reasonable if no action was taken.
- **None:** No commitment is expressed.

	A & B κ		*A & C κ*		*B & C κ*		*3-Way κ*	
Message Types	B	S	B	S	B	S	B	S
All	0.83	0.57	0.79	0.66	0.74	0.56	0.78	0.60
Single Recipient	0.85	0.54	0.79	0.66	0.80	0.55	**0.81**	0.58
Closed Group	0.79	0.55	0.79	0.58	0.66	0.61	0.75	0.58
Broadcast	0.82	0.66	0.77	0.70	0.70	0.53	0.76	0.63

Table B.1: Pairwise and three-way κ agreements for classifying requests in the *Sentence 1* annotation experiment *(B=Binary Agreement, S=Fine-grained Strength Agreement)*.

	A & B κ		A & C κ		B & C κ		3-Way κ	
Message Types	B	S	B	S	B	S	B	S
All	0.45	0.30	0.62	0.44	0.55	0.37	0.54	0.37
Single Recipient	0.51	0.28	0.76	0.54	0.52	0.41	**0.60**	0.41
Closed Group	0.14	0.06	0.51	0.32	0.41	0.19	0.35	0.19
Broadcast	0.52	0.41	0.52	0.40	0.66	0.41	0.57	0.41

Table B.2: Pairwise and three-way κ agreements for classifying commitments in the *Sentence 1* annotation experiment *(B=Binary Agreement, S=Fine-grained Strength Agreement)*.

B.1.5 Audience Types

For each email message in our pilot annotation corpus, we also manually classified the nature of the recipient audience. Each email was classified with one of the following message types:

- **Single Recipient:** Addressed to a single recipient. We also consider email messages that are addressed specifically to a single recipient but carbon copied or blind carbon copied to another recipient to belong to this group.

- **Closed Group:** Addressed explicitly to a specified group of recipients. Each recipient must be identifiable from the email headers.

- **Broadcast:** Addressed to an unspecified group of recipients, such as a group alias or mailing list.

We used information about the audience type of email messages in analysing inter-annotator agreements for our annotation task, as we describe below in Section B.1.6.

B.1.6 Results and Discussion

The inter-annotator agreement for our pilot annotation are shown in Table B.1 and Table B.2. Note that all κ values referred to in this section, apart from the pairwise κ values in Tables B.1 and B.2, refer to agreement between all three annotators calculated using the generalisation of Cohen's Kappa to more than two annotators, as specified by Krippendorff (1980).

The results in Table B.1 show pairwise and three-way inter-annotator agreement for marking requests in our pilot annotation task. Separate BINARY and STRENGTH κ values are given for each measurement.

Binary agreement refers to inter-annotator agreement about which sentences contain a request, ignoring any indication of strength. To calculate these κ scores, we collapsed all three strengths of annotated requests into a single request class. Thus, disagreement about the strength of a request is ignored in binary κ scores.

Strength agreement, which is always lower than binary agreement, refers to inter-annotator agreement for the more fine-grained strength categories of requests. Thus, it represents agreement between annotators over which sentences contain a STRONG, MEDIUM, WEAK or NO REQUEST. Disagreement about the strength of an identified request is considered a complete disagreement for the strength κ scores.

The results in Table B.2 similarly show both pairwise and three-way inter-annotator agreement for classifying commitments. The same separate measures of binary and strength agreement are used.

Finally, both Table B.1 and Table B.2 also show separate κ scores for subsets of the annotation corpus, grouped according to the email audience type (see Section B.1.5).

From these results, we can draw several tentative conclusions about human agreement for identifying requests and commitments in email:

1. There is good agreement ($\kappa = 0.78$) about which sentences embody a request.

2. There is some tentative agreement ($\kappa = 0.60$) about the strength of requests.

3. There is poorer agreement about which sentences embody a commitment ($\kappa = 0.54$) and poor agreement about the strength of those commitments ($\kappa = 0.37$).

4. The level of agreement appears to vary depending upon the audience type of the email message.

A particularly interesting aspect of our results is the variation in inter-annotator agreements across the different audience types that we identified in Section B.1.5. As can be seen both in Tables B.1 and B.2, agreement between our three annotators about the presence and strength of both requests and commitments is highest for Single Recipient email messages, and lowest for Closed Group email messages. In the case of commitments, this difference is particularly marked. Analysing the cases of disagreement did not revealed a recurring reason for these differences. Some annotators did make observations that the strength of requests may be determined, to some extent, by the probability that it would apply to the specified recipient when an email message has multiple recipients. It is unclear how to objectively judge such a probability, but perhaps observations such as this shed some light on the reduced agreement for Closed Group email messages.

In addition to the annotations made regarding requests and commitments, we allowed users to flag segmentation errors, as noted in Section B.1.2. Although we do not consider this method a rigorous way to evaluate the SentParBreak sentence splitter, it is interesting to note that the results from this error flagging suggest that segmentation error for our pilot annotation data was at least 10%. This is a much greater error rate than the 0.997352 precision and 0.995093 recall results that were apparently achieved using SentParBreak over the Genia corpus (Piao, 2007). Such a difference in performance served to highlight the need for standard NLP tools like sentence segmenters to be retrained, retuned or otherwise tailored when working with email data, due to differences in the nature of textual content.

B.2 Sentence 2: A Second Sentence-Level Annotation Experiment

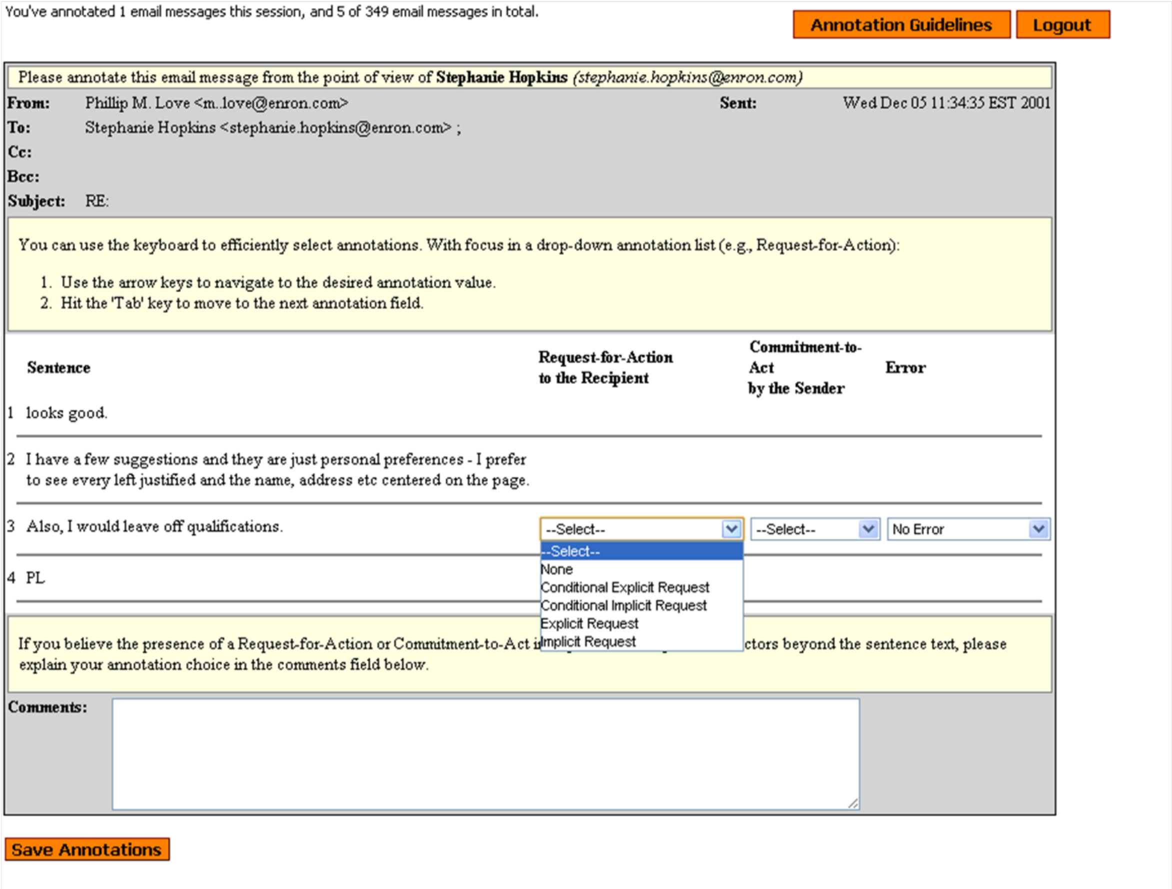
Following on from the pilot sentence annotation task described in Section B.1, we performed a second sentence-level annotation task. This second annotation task used the same approach as the first, with refinements to:

- The guidelines provided for recognising requests and commitments. We provided annotators with more explicit guidance for marking conditional and implicit requests and commitments.
- The fine-grained categories for annotation. Rather than marking strength, which we found weak agreement for in our pilot annotation task, we asked annotators to explicitly identify conditionality and explicitness of the requests and commitments they marked.
- The annotation tool. We modified the annotation tool to support the different annotation categories required, as shown in Figure B.2.

B.2.1 Annotation Guidelines

The annotation guidelines included modifications to the first set of guidelines focused on the inclusion of information about explicitness and conditionality.

The revised guidelines noted that both requests and commitments may be conditional, in the sense that scheduling or executing the specified action is expected only if some stated condition is satisfied. For both requests and commitments, the condition is usually stated in the utterance containing the request or commitment. It may, however, be stated elsewhere in the email message, and understood to apply to the request or commitment utterance, as in the following example:



You've annotated 1 email messages this session, and 5 of 349 email messages in total.

Please annotate this email message from the point of view of **Stephanie Hopkins** *(stephanie.hopkins@enron.com)*

From: Phillip M. Love <m..love@enron.com> **Sent:** Wed Dec 05 11:34:35 EST 2001
To: Stephanie Hopkins <stephanie.hopkins@enron.com> ;
Cc:
Bcc:
Subject: RE:

You can use the keyboard to efficiently select annotations. With focus in a drop-down annotation list (e.g., Request-for-Action):

1. Use the arrow keys to navigate to the desired annotation value.
2. Hit the 'Tab' key to move to the next annotation field.

	Sentence	Request-for-Action to the Recipient	Commitment-to-Act by the Sender	Error
1	looks good.			
2	I have a few suggestions and they are just personal preferences - I prefer to see every left justified and the name, address etc centered on the page.			
3	Also, I would leave off qualifications.	--Select--	--Select--	No Error
4	PL			

If you believe the presence of a Request-for-Action or Commitment-to-Act i... ...ctors beyond the sentence text, please explain your annotation choice in the comments field below.

Comments:

Figure B.2: The modified web-based annotation tool used in the *Sentence 2* annotation task.

(B.2) *Are you coming to the meeting? If so, send me items for the agenda by 2pm.*

Note also that conditional commitment binds the relevant party to action only when the stated condition is satisfied:

(B.3) *I'll send a draft if Harry responds before Friday.*

Direct requests state the actual request explicitly, either in the form of an imperative command, as in:

(B.4) *Please attend the meeting tomorrow.*

or as a question where a response to the literal interpretation of the question would result in a speech act with the required content, as in:

(B.5) *What time does your flight arrive tomorrow?*

	Conditional	Unconditional
Explicit	*Please complete the project plan by Thursday if you want comments*	*Please complete the project plan by Thursday*
Implicit	*If you want comments, I need the project plan by Thursday.*	*I need the project plan by Thursday.*

Table B.3: Minimal pair examples of requests, highlighting conditionality and explicitness, as defined for the *Sentence 2* annotation experiment.

Note that the content may be information, permission, interpretation and so on, as noted in our definition.

Indirect requests may or may not state the required action, but do not explicitly instruct the recipient to act or respond with the required speech act. Thus, Examples (B.6) and (B.7) are direct requests, while Example (B.8) is indirect, since a literal response would be a yes/no answer.

(B.6) *Please send me my curves and trades for Jan 18.*

(B.7) *What were they?*

(B.8) *Can you send my curves and trades for Jan 18?*

Similarly, direct commitments state the promise for future action explicitly. Indirect commitments do not state the promise to complete future action. Note that in either case, as for requests, the actual action may or may not be stated explicitly. So, Examples (B.9) and (B.10) represent direct commitments, while Example (B.11) is considered indirect.

(B.9) *I'll send you the document today.*

(B.10) *I'll do it.*

(B.11) *Leave it to me.*

In the annotation guidelines, we also provided tables of minimal pair examples to help clarify conditionality and explicitness of requests and commitments. These tables, drawn directly from the annotation guidelines, are shown in Table B.3 and B.4.

B.2.1.1 Results and Discussion

Agreements for our second experiment (using the same three human annotators as our first annotation task) are summarised in Table B.5. We discuss these results further in Section 3.2.3.

	Conditional	Unconditional
Explicit	*If you need me to, I'll book a room*	*I'll book a room.*
Implicit	*Let me know if you need a room.*	*Consider the room booked.*

Table B.4: Minimal pair examples of commitments, highlighting conditionality and explicitness, as defined for the *Sentence 2* annotation experiment.

	A & B κ			A & C κ			B & C κ			3-Way κ		
	B	C	E	B	C	E	B	C	E	B	C	E
Request	0.85	0.82	0.57	0.63	0.60	0.52	0.73	0.53	0.71	0.74	0.65	0.60
Commit	0.87	0.80	0.83	0.74	0.79	0.69	0.79	0.73	0.61	0.80	0.77	0.71

Table B.5: Pairwise and three-way κ agreements for classifying requests and commitments in the *Sentence 2* annotation experiment *(B=Binary Agreement, C=Fine-grained Conditionality Agreement, E=Fine-grained Explicitness Agreement).*

B.3 Message 1: An Initial Message-Level Annotation Experiment

This annotation experiment was conducted to explore the effect of the different unit size on human agreement.Specifically, we were interested in the effect of moving from sentence-level annotation to message-level annotation. As expected, inter-annotator agreement for both requests and commitments increased over the sentence-level experiments.

This experiment was conducted with a different set of 100 messages, again drawn from the Enron email dataset. Three annotators marked these messages independently, using another modified version of our web-based annotation tool. This time, annotators indicated binary annotations at the message level, without regard for specifically indicating strength, explicitness or conditionality.

B.3.1 Annotation Guidelines

Annotators were provided with the following annotation guidelines.

Thank you for participating. You will be presented with a series of email messages, and should annotate each message as follows:

First indicate whether it expresses a request. Second indicate whether it expresses a commitment. Finally, record reasoning for your annotation decision, or any other issues in the Comments field. Definitions of request and commitment are provided below.

When annotating, please:

- Annotate based on the content written by the current email sender. Any quoted or forwarded content that may be present should only be used to help interpret or clarify the intent of new content in an email message.

- Annotate from the point of view of the specified recipient for each email message. The specified recipient is noted above each email message.
- Provide explanation for your annotation in the Comments field if you believe the presence of a request or commitment is context-dependent (e.g., on the relationship between sender and receiver).
- You may stop annotating and logout at any time. All submitted annotations will be saved. You can return and continue annotating at any time. Once you have read and understood these instructions, please begin.

Choosing the Right Annotation

Annotate messages independently for both requests and commitments, and note that a single message may contain both a request and a commitment.

Requests

Mark as a request any email message that requires the recipient to execute or schedule an action or to respond.

The following further define utterances that are considered to be requests:

1. A request is an utterance from the email sender that places an obligation on the email recipient to:

   - schedule an action (e.g., by adding something to a calendar or task list);
   - perform (or not perform) an action; or
   - respond with some speech act.

   Examples:

   (B.12) *Please come along to the meeting next Friday at 2pm.*

   (B.13) *Make sure you finish the report today.*

   (B.14) *Are you in the office tomorrow?*

2. A speech act response to a request may need to be directed to the email sender or to another party. Speech act responses are not limited to any specific type of speech act, but often include informing, permitting, committing, accepting and refusing.

Examples:

(B.15) *What time are we meeting today?*

(B.16) *Let John know whether you'll be attending.*

(B.17) *Is it ok if I come in late on Friday?*

(B.18) *Please tell me when you'll be in Sydney.*

3. Conditional requests oblige action or response from the email recipient only if some stated condition is satisfied. The stated condition is often in the same utterance as the request, but may be elsewhere in the email message.

Examples:

(B.19) *If you want comments, I need the project plan completed today.*

(B.20) *You'll need to finish the plan today if you want my feedback.*

(B.21) *To order a new laboratory notebook, talk to Andy in the store.*

4. Unconditional requests require action or response without consideration of any preconditions.

Examples:

(B.22) *Please come along to the meeting next Friday at 2pm.*

(B.23) *Can you please send me the plan?*

5. Direct requests are surface-form realisations which express the request for action or response explicitly, usually as an imperative command or a question where a response to the literal interpretation of the request would result in a speech act with the required information or content.

Examples:

(B.24) *Please send me the project plan today.*

(B.25) *What time is the meeting today?*

6. Indirect requests are surface-form realisations that do not explicitly instruct the recipient to act or to respond with the required speech act. For scheduling or executing action, the requested action or speech act itself may be stated, but not the instruction to act or respond.

Examples:

(B.26) *I need you to send me the project plan today.*

(B.27) *Do you know what time the meeting is today?*

(B.28) *I was wondering if you might be able to introduce me to John.*

7. Requests for inaction are considered to be requests.

Examples:

(B.29) *Don't discuss this project with anyone outside the organisation.*

The following are not considered to be requests:

1. Pleasantries are utterances that resemble a request, but place little or no obligation on the recipient to respond. Pleasantries are not considered to be requests. Their presence is largely formulaic, and their function is primarily to conform to social conventions rather than to elicit action or response.

   Variations on *Let me know if you have any questions* are particularly common in email messages. The context of the entire email message needs to be considered in order to distinguish between when such a statement functions as an actual request for action or response, and when it is merely a pleasantry.

Examples:

(B.30) *How are you?*

(B.31) *Let me know if you have any questions.*

Commitments

Mark as a commitment any message that promises to execute or schedule a future action or response. The agent obliged to act or respond may be the sender or an identifiable third-party.

The following further define utterances that are considered to be commitments:

1. A commitment is any utterance in an email message written by the sender that promises or offers future action by someone other than the email recipient. In particular, the future action should be something that the email recipient might be waiting for someone to do or cares about and might want to follow up with the responsible person. Given that a commitment refers to future action, you might expect to find the action on the responsible person's task list or calendar.

Examples:

(B.32) *I'll be at the meeting next Friday at 2pm.*

(B.33) *I can book a meeting room if you like.*

(B.34) *Please let me know where you are located, and I'll try to get it over there today.*

2. The obligation to complete the promised or offered action may be on the email sender or some other person, group of people, or organisation (not including the email recipient).

Examples:

(B.35) *I will do my best to finish the report today.*

(B.36) *John will send through the report as soon as it's finished.*

(B.37) *Our finance department will pay the invoice by the end of the month.*

3. Conditional commitments only commit to future action if some stated condition is satisfied. The stated condition is often in the same utterance as the commitment, but may be elsewhere in the email message.

Examples:

(B.38) *If you wish, I could provide you questions in advance.*

(B.39) *Katherine will give you the report if I'm not at my desk.*

4. Unconditional commitments commit to future action without consideration of any preconditions.

Examples:

(B.40) *Either Sean or I will forward the e-mails to you.*

(B.41) *I'll keep you posted on any changes.*

5. Direct commitments are surface-form realisations that state the commitment to future action explicitly. They may or may not state the future action explicitly.

Examples:

(B.42) *I'll book a room.*

(B.43) *I'll do it.*

(B.44) *John will book a room if we need one.*

6. Indirect commitments are surface-form realisations that do not explicitly state the commitment to future action. They may or may not state the future action.

Examples:

(B.45) *More details to follow.*

(B.46) *John can book a room.*

(B.47) *Leave the room arrangements to me.*

(B.48) *Leave it to me.*

7. Commitments to inaction are considered to be commitments.

Examples:

(B.49) *I won't send any further updates.*

(B.50) *John won't be in the office on Friday.*

The following are not considered to be commitments:

1. If the party committed to the future action is not identifiable, the utterance is not considered to be a commitment.

Examples:

(B.51) *The briefing will be in Room F on the first floor.*

2. Pleasantries are socially-oriented utterances that resemble the form of a commitment, but place little or no obligation on the sender (or other assigned party) to act. Pleasantries are not considered to be commitments. Their presence is largely formulaic, and their function is not primarily to commit to any future action or response.

   Variants on *Let me know if you have any questions* are particularly common in email messages. The context of the entire email message is considered in order to distinguish between when such a statement is present due to social convention and when it functions as an actual offer or commitment for future action or response, usually from the sender.

Examples:

(B.52) *Let me know if you have any questions.*

(B.53) *Please call if you require any further information.*

Type	3-Way κ
Request	0.84
Commitment	0.78

Table B.6: Three-way κ agreements for classifying requests and commitments in the *Message 1* annotation experiment.

Type	5-Way κ	3-Way κ[2]	Best 3-Way κ [3]
Request	0.66	0.70	0.79
Commitment	0.66	0.66	0.80
Pleasantry	0.71	0.70	0.77

Table B.7: Three-way κ agreements for classifying requests, commitments and pleasantries in the *Message 2* annotation experiment.

B.3.2 Inter-Annotator Agreement

Across the 100 message corpus we observed good agreement for both requests (κ=0.84) and commitments (κ=0.78), as shown in Table B.6. The annotators were the same three people who originally completed both the 'Sentence 1' and 'Sentence 2' annotation experiments, so had no doubt benefited from discussion and consideration of the issues that arose from analysis of the disagreements in those experiments.

B.4 Message 2: Message-Level Annotation with Five Annotators

This experiment used the same guidelines as the 'Message 1' experiment, and expanded the pool of annotators to five people, two of whom had no introduction to the problem of identifying requests and commitments before they started the annotation task. The only instructions they received were the set of annotation guidelines included above in Section B.3.1. The corpus annotated in this experiment was a separate set of 209 email messages randomly extracted from the Enron email corpus with no human filtering.

B.4.1 Inter-Annotator Agreement

The agreement levels from this experiment were mixed. We found the best agreement was not actually between the three experienced annotators, but between two experience and one novice annotator. The main cause of disagreement among our experienced annotators stemmed from differing interpretations of requests related to attachments, and commitments related to meetings.

Type	A + B κ	B + C κ	A + C κ	3-way κ
Request	0.81	0.87	0.88	0.85
Commitment	0.78	0.86	0.93	0.86
Pleasantry	0.48	0.69	0.80	0.66

Table B.8: Pairwise and three-way κ agreements for classifying requests, commitments and pleasantries in the *Message 3* annotation experiment.

B.5 Message 3: Expanding Message-Level Annotation

To consolidate the data gathered in the *Message 1* and *Message 2* experiments, we conducted an annotation experiment with a larger set of email messages from the Enron dataset. This resulted in a collection of 664 annotated messages that had been selected at random from the full Enron dataset.

Each message was annotated by three annotators, with overall kappa agreement of 0.681.The unanimously agreed data set consists of 505 email messages.

B.5.1 Inter-Annotator Agreement

The inter-annotator agreements for the *Message 3* experiment are shown in Table B.8. The average agreement across the three annotators was 0.85 for requests was 0.86 for commitments, which represents better message-level agreement than was achieved in any of our other, smaller-scale annotation experiments. Notably, agreement on pleasantries, as we were still calling phatic requests and commitments during this experiment, was significantly lower, with a three-way κ agreement of 0.66.

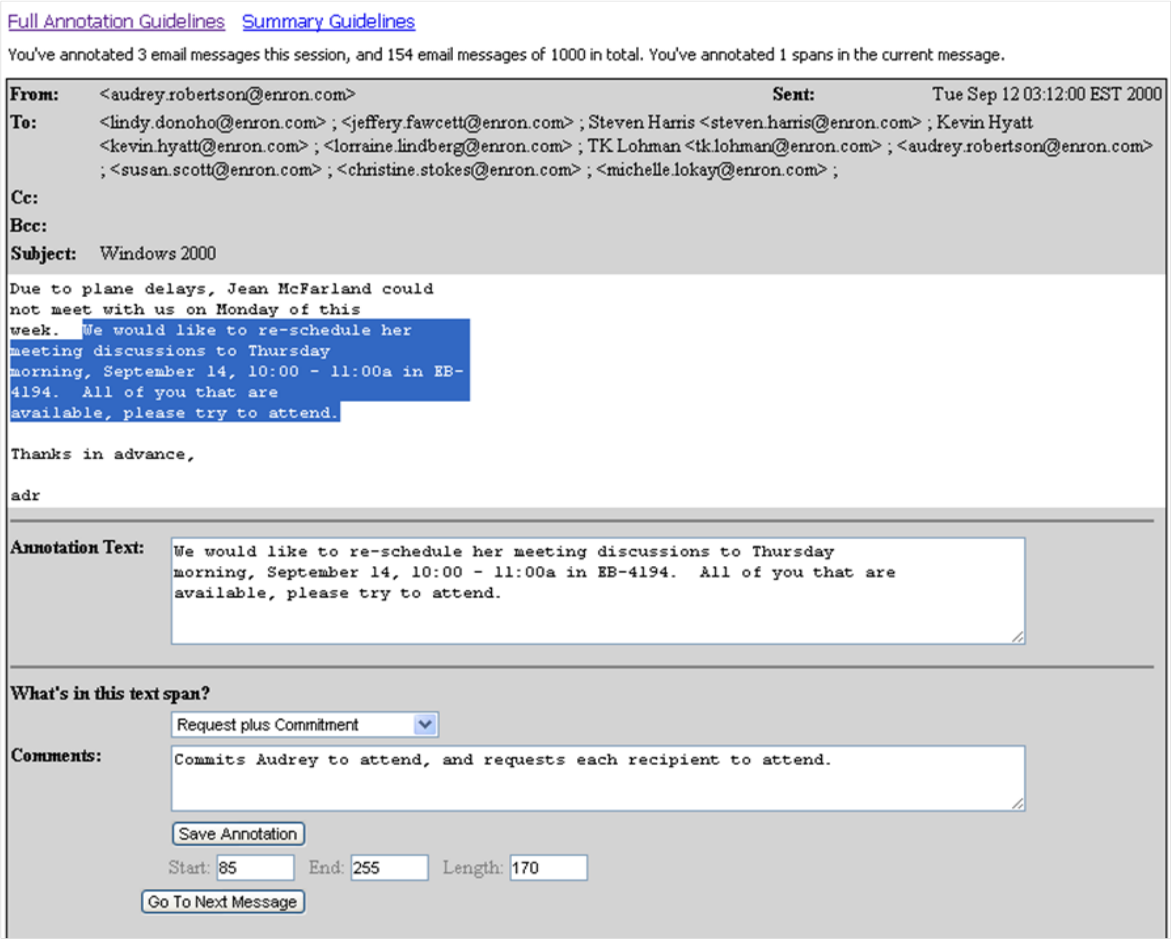
B.6 Span 1: A Pilot Span Annotation Experiment

As requests and commitments are not realised with fixed-sized units of text, we undertook this experiment as our first, pilot free text annotation task that required annotators to both identify the extent of text they wished to annotate, and then choose the appropriate label. We refer to this tasks as a SPAN ANNOTATION tasks.

This was the first of our annotation experiments that did not fix the unit of annotation. Our interest in requiring annotators to identify both the extent of the request

[2] This measures three-way agreement for the annotators involved in the Sentence 1, Sentence 2 and Message 1 experiments.

[3] This measures the best three-way agreement between triples of annotators selected from our five annotators.



You've annotated 3 email messages this session, and 154 email messages of 1000 in total. You've annotated 1 spans in the current message.

From: <audrey.robertson@enron.com> **Sent:** Tue Sep 12 03:12:00 EST 2000

To: <lindy.donoho@enron.com> ; <jeffery.fawcett@enron.com> ; Steven Harris <steven.harris@enron.com> ; Kevin Hyatt <kevin.hyatt@enron.com> ; <lorraine.lindberg@enron.com> ; TK Lohman <tk.lohman@enron.com> ; <audrey.robertson@enron.com> ; <susan.scott@enron.com> ; <christine.stokes@enron.com> ; <michelle.lokay@enron.com> ;

Cc:

Bcc:

Subject: Windows 2000

```
Due to plane delays, Jean McFarland could
not meet with us on Monday of this
week.  We would like to re-schedule her
meeting discussions to Thursday
morning, September 14, 10:00 - 11:00a in EB-
4194.  All of you that are
available, please try to attend.

Thanks in advance,

adr
```

Annotation Text: We would like to re-schedule her meeting discussions to Thursday morning, September 14, 10:00 - 11:00a in EB-4194. All of you that are available, please try to attend.

What's in this text span? Request plus Commitment

Comments: Commits Audrey to attend, and requests each recipient to attend.

Start: 85 End: 255 Length: 170

Figure B.3: A screenshot of our web-based span-annotation tool, built to gather fine-grained annotations on the specific location of requests and commitments in the body text and subject lines of email messages. The user selects the relevant text in the mail message, and the relevant fields in the bottom half of the window are automatically populated.

or commitment as well as its label was two-fold:

1. We wanted to explore the nature of text extents that were used to realise requests and commitments;

2. We wanted to generate annotated email data that could be used to train fine-grained request and commitment classifiers. Using span-level data gave us maximum flexibility to be able to generate fine-grained training corpora at a range of different granularities, as detailed in Section 5.1;

Our review of related literature also suggested that no such corpus had previously been made publicly available. We therefore saw value in creating our own labelled corpus of span-level requests and commitments that could be publicly released for use in other email research.

This tasks was performed using another custom-built annotation tool, shown in Figure B.3. This tool allows annotators to mark textual units of any size, from a single character up to the entire content of an email message, and then to label the span as a request, a commitment, a phatic obligation act, or combinations of these. As shown in Figure B.3, an annotator first highlights a span of text in the message body or subject, and then selects a label from the closed set of available annotations.

The available annotations that can be used to label each span, are:

- Request Only;
- Commitment Only;
- Request plus Commitment;
- Phatic Request;
- Phatic Commitment;
- Request plus Phatic Commitment;
- Commitment plus Phatic Request;
- Phatic Request plus Phatic Commitment; and
- None.

It may seem unexpected that we include a 'None' category label, as we would expect text spans that do not represent either a request or commitment to simply not be identified. Certainly, the vast majority of non-obligation act text spans are not marked. We explicitly include a 'none' category, however, to allow our annotators to remove ambiguity about utterances for which they considered their interpretation as a non-obligation act subjective, ambiguous or otherwise controversial. In this way, annotators could explicitly signal a negative interpretation. In contrast, the absence of annotation does not rule out possible oversight of the text span, rather than a negative interpretation. In practice, 62 spans were labelled with a 'None' label, including utterances that represented rhetorical questions and directive utterances in unsolicited commercial email that were not considered actionable.

Annotators were also instructed to add a free text comment in the event of any ambiguity, uncertainty or context-sensitivity. This process can be repeated as often as necessary to allow each annotator to mark all request and/or commitment spans within each message before moving to the next message.

In this *Span 1* experiment, three independent annotators used the span annotation tool to mark all request and commitment spans in a collection of 50 business email messages, drawn from the Enron email corpus. The annotation guidelines which annotators were given are provided in Appendix A.1. This task was used largely to identify any flaws in our annotation process or tools. As no significant short-comings were identified, and our annotation guidelines remained consistent across our *Span 1* and

Annotators	Request F-score	Commitment F-score
A + B	0.949	0.850
B + C	0.882	0.845
A + C	0.906	0.896
Average	0.912	0.864

Table B.9: Pairwise and three-way F-score agreements for annotating request and commitment spans in the *Span 1* Pilot annotation task.

Span 2 experiments, we were also able to use the annotated data from this experiment as a separate corpus of test data for our fine-grained classification experiments in Chapter 5.

B.6.1 Inter-Annotator Agreement

Measuring agreement for span annotation is different to our previous annotation experiments, since it cannot easily be interpreted as a text classification task. This makes inter-annotator agreement measures like Cohen's kappa inapplicable for measuring agreement on the task of text-span agreement. Instead, a brief review of the literature suggests that F-score is frequently used in such cases, where the unit of text annotation is not fixed. One way to approach calculating precision and recall between two annotators is to treat the annotations of one annotator as the 'gold standard' and calculate precision and recall for the other annotator against this reference. In practice, it makes no difference (to F-score) which annotator is selected as the gold-standard (precision and recall are simply inverted between annotators).

Calculating these results, using an agreement metric that allows a fifteen character-window threshold for text span start character and overall text span length differences, gives us the following results:

B.7 Span 2: A Large-scale Span Annotation Experiment

The resulting annotated corpus contains 3397 labelled spans of text. The longest annotated span was 2198 characters long, containing a sequence of instructions to follow in the event of receiving a package suspected of containing Anthrax, marked as a request; the shortest agreed span was 7 characters, consisting of the utterance *cancel?*, in response to an interview request email, which was marked as a request by both annotators.

This experiment involved two annotators marking all request and commitment spans within 1000 email messages drawn from the Enron corpus.

Obligation Act	Distinct Spans	F-score
Request	2318	0.855
Commitment	644	0.723

Table B.10: Pairwise F-score agreements for annotating request and commitment spans in the *Span 2* annotation task.

As for the *Span 1* experiment, this task was performed using the custom-built annotation tool, shown in Figure B.3.

The annotation guidelines for our *Span 2* experiment were the same as those used for the *Span 1* experiment, and are included in Appendix A.

B.7.1 Inter-Annotator Agreement

We measure agreement using the same F-score mechanism that we employed in the *Span 1* experiment. As shown in Table B.10, we achieve an F-score agreement of 0.86 for requests and 0.72 for commitments, using a word-overlap span agreement metric with a minimum threshold of one word.

Bibliography

Akbani, Rehan, Stephen Kwek, and Nathalie Japkowicz. 2004. Applying Support Vector Machines to Imbalanced Datasets. In *Proceedings of the 15th European Conference on Machine Learning (ECML)*, pages 39–50, Pisa, Italy, September 20–24.

Alexandersson, Jan, Marc Al-Hames, Mihaela Bobeica, Jan Cernocky, Alfred Dielmann, Michal Fapso, Daniel Gatica-Perez, Yoshi Gotoh, Sabrina Hsueh, Natasa Jovanovic, Thomas Kleinbauer, Wessel Kraaij, Stephan Lesch, Harald Lochert, Johanna Moore, Gabriel Murray, Stephan Reiter, Steve Renals, Ruther Rienks, Gerhard Rigoll, Petr Schwarz, Stanislav Sumec, Weiqun Xu, and Dong Zhang. 2005. Report on Initial Work in Segmentation, Structuring, Indexing and Summarization. Technical Report FP6-506811, Augmented Multiparty Interaction.

Alexandersson, Jan, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. Dialogue acts in VERBMOBIL-2. Technical Report Verbmobil-Report 226, DFKI Saarbruecken, Universitt Stuttgart, Technische Universitt Berlin, Universitt des Saarlandes. 2nd Edition.

Alexandersson, Jan, Ralf Engel, Michael Kipp, Stephan Koch, Uwe Küssner, Norbert Reithinger, and Manfred Stede. 2000. Modeling Negotiation Dialogs. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, pages 441–451.

Apache Foundation, The. 2013. Apache OpenNLP. Apache Foundation. http://opennlp.apache.org/. Accessed: 28 April 2013.

Archer, Dawn, Jonathan Culpeper, and Matthew Davies. 2008. Pragmatic annotation. In Anke Ludeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, pages 613–642.

Austin, John L. 1962. *How to do things with words*. Harvard University Press.

Bach, Kent and Robert M Harnish. 1979. *Linguistic Communication and Speech Acts*. The MIT Press.

Baron, Jason R, David D Lewis, and Doug W Oard. 2006. Trec-2006 legal track overview. In *The Fifteenth Text REtrieval Conference (TREC) Proceedings*, pages 79–98, Gaithersburg, Maryland, USA, November 14–17.

Baron, Naomi S. 1998. Letters by phone or speech by other means: the linguistics of email. *Language and Communication*, 18(2):133–170, April.

Beeferman, Douglas, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 35–46, Providence, RI, USA, August 1–2.

Beigman, Eyal and Beata Beigman Klebanov. 2009. Learning with Annotation Noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP) of the Asian Federation of Natural Language Processing (AFNLP)*, pages 280–287, Singapore, August 2–9.

Bellotti, Victoria, Nicolas Ducheneaut, Mark Howard, and Ian Smith. 2003. Taking Email To Task: The Design and Evaluation of a Task Management Centred Email Tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 345–352, Ft Lauderdale, Florida, USA, April 5–10.

Bennett, Paul N and Jaime G Carbonell. 2005. Feature representation for effective action-item detection. In *Proceedings of Beyond Bag-of-Words Workshop at the 28th Annual International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pages 63–70, Salvador, Brasil, August 15–19.

Bennett, Paul N and Jaime G Carbonell. 2007. Combining Probability-Based Rankers for Action-Item Detection. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 324–331, Rochester, NY, April 22–27.

Bilbow, Grahame T. 2002. Commissive speech act use in intercultural business meetings. *International Review of Applied Linguistics in Language Teaching*, 40(4):287–303, November.

Brown, Steven D., David Middleton, and Geoffrey Lightfoot. 2001. Performing the Past in Electronic Archives: Interdependencies in the Discursive and Non-Discursive Ordering of Institutional Rememberings. *Culture and Psychology*, 7(2):123–144, June.

Bunt, Harry. 1994. Context and Dialogue Control. *THINK Quarterly*, 3(1):19–31, May.

Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2548–2555, Valletta, Malta, May 19–21.

Camino, Beatrice M., Allen E. Milewski, David R. Millen, and Thomas M. Smith. 1998. Replying to email with structured responses. *International Journal of Human-Computer Studies*, 48(6):763–776, June.

Carasik, Robert P. and Charles E. Grantham. 1988. A case study of cscw in a dispersed organisation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 61–66, Washington D.C., USA, May 15–19.

Carvalho, Vitor R and William W Cohen. 2004. Learning to Extract Signature Reply Lines from Email. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, USA, July 30–31.

Carvalho, Vitor R. and William W. Cohen. 2006. Improving Email Speech Act Analysis via N-gram Selection. In *Proceedings of Workshop on Analyzing Conversations in Text and Speech (ACTS) at Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 35–41, New York, USA, June 8–9.

Carvalho, Vitor R.; Cohen, William W;. 2005. On the Collective Classification of Email 'Speech Acts'. In *Proceedings of the 28th Annual International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, Salvador, Brazil, August 15–19.

Cavender, Sasha. 1998. Legends. Forbes, May. http://www.forbes.com/asap/1998/1005/126.html.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, January.

Chen, Hao, Jianying Hu, and Richard W Sproat. 1999. Integrating geometrical and linguistic analysis for email signature block parsing. *ACM Transactions on Information Systems*, 17(4):343–366, October.

Clark, Alexander and Andrei Popescu-Belis. 2004. Multi-level Dialogue Act Tags. In *5th SIGdial Workshop on Discourse and Dialogue*, pages 163–170, Cambridge, MA, USA, April 30 – May 1.

Clark, Herbert H. 1997. *Using Language*. University Press, Cambridge.

Cohen, William W., Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into 'Speech Acts'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 309–316, Barcelona, Spain, Jul 25–26.

Conklin, Jeff and Michael L. Begeman. 1988. gIBIS: a hypertext tool for exploratory policy discussion. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW)*, pages 140–152, Portland, OR, USA.

Core, Mark and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA, USA, November 8–10.

Corston-Oliver, Simon H., Eric Ringger, Michael Gamon, and Richard Campbell. 2004. Task-focused summarization of email. In *Proceedings of the Text Summarization Branches Out Workshop at the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 43–50, Barcelona, Spain, July 22–25.

Crystal, David. 2001. *Language and the Internet*. Cambridge University Press, Cambridge, UK.

Culotta, Aron, Ron Bekkerman, and Andrew McCallum. 2004. Extracting Social Networks and Contact Information from Email and the Web. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, USA, July 30–31.

Dabbish, Laura A., Robert E. Kraut, Susan Fussell, and Sara Kiesler. 2005. Understanding Email Use: Predicting Action on a Message. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, Portland, OR, USA, April 2–7.

Daly, Bill. 1996. Electronic mail: strangely familiar texts. Originally available at http://cougar.vut.edu.au/~dalbj/e-mail.htm.

Danet, Brenda. 1998. Computer-mediated communication. In Paul Bouissac, editor, *Encyclopedia of Semiotics*. Oxford University Press, New York, chapter Computer-mediated Communication.

De Felice, Rachele and Paul Deane. 2012. Identifying speech acts in e-mails: Toward automated scoring of the TOEIC e-mail task. ETS Research Report RR-12-16, Educational Testing Service (ETS), Princeton, NJ, USA, September.

Denning, Peter J. 1982. Acm president's letter: Electronic junk. *Communications of the ACM*, 25(3):163–165.

Ducheneaut, Nicolas and Victoria Bellotti. 2001. E-mail as habitat: an exploration of embedded personal information management. *Interactions*, 8(5):30–38, September/October.

Estival, Dominique, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author Profiling for English Emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 263–272, Melbourne, Australia, September 19–21.

Fallows, Deborah. 2002. E-mail at work: few feel overwhelmed and most are pleased with the way e-mail helps them do their jobs. Technical report, Pew Internet and American Life Project.

Fallows, Deborah. 2007. Spam 2007. Technical report, Pew Internet and American Life Project.

Faulring, Andrew, Ken Mohnkern, Aaron Steinfeld, and Brad A. Myers. 2009. Design and Evaluation of User Interfaces for the RADAR Learning Personal Assistant. *AI Magazine*, Winter:74–84.

Faulring, Andrew, Brad Myers, Ken Mohnkern, Bradley Schmerl, Aaron Steinfeld, John Zimmerman, Asim Smailagic, Jeffery Hansen, and Daniel Siewiorek. 2010. Agent-Assisted Task Management that Reduces Email Overload. In *Proceedings on Intelligent User Interfaces (IUI)*, pages 61–70, Hong Kong, China, February 7–10.

Ferguson, Paul, Neil OH́are, Michael Davy, Adam Bermingham, Paraic Sheridan, Cathal Gurrin, and Alan F. Smeaton. 2009. Exploring the use of Paragraph-level Annotations for Sentiment Analysis of Financial Blogs. In *Proceedings of the Workshop on Opinion Mining and Sentiment Analysis (WOMSA) at the Conference of the Spanish Association for Artificial Intelligence (CAEPIA-TTIA)*, pages 42–52, Seville, Spain, 13 November.

Fernandez, Raul and Rosalind W Picard. 2002. Dialog act classification from prosodic features using support vector machines. In *Proceedings of Speech Prosody*, pages 291–294, Aix-en-Provence, France, April 11–13.

Fisher, Danyel, A J Brush, E Gleave, and Mark A Smith. 2006. Revisiting Whittaker and Sidner's "E-mail Overload" ten years later. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 309–312, Banff, Alberta, Canada, November 4–8.

Franzén, Kristofer, Gunnar Eriksson, Frederik Olsson, Larse Asker, Per Lidén, and Joakim Cöster. 2002. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1–3):49–61, December.

Freed, Michael, Jaime Carbonell, Geoff Gordon, Jordan Hayes, Brad Myers, Daniel Siewiorek, Stephen Smith, Aaron Steinfeld, and Anthony Tomasic. 2008. RADAR: A Personal Assistant that Learns to Reduce Email Overload. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence (AAAI)*, pages 1287–1293, Chicago, USA, July 13–17.

Freed, Ned and Nathaniel Borenstein. 1996. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. RFC 2045, Internet Engineering Task Force (IETF), November.

Gellens, Randall. 2004. The Text/Plain Format and DelSp Parameters. RFC 3676, Internet Engineering Task Force (IETF), February.

Goldspink, Chris. 2007. Normative self-regulation in the emergence of global network institutions: the case of Wikipedia. In *Proceedings of Australia New Zealand Systems Conference*, Auckland, NZ, December 2–5.

Goldstein, Jade and Roberta Evans Sabin. 2006. Using speech acts to categorize email and identify email genres. In *Proceedings of the 39th Hawaii International Conference on System Sciences (HICSS)*, page 50b, Kauai, Hawaii, USA, January 4–7.

Gwizdka, Jacek. 2002. Future Time in Email - Design and Evaluation of a Task-based Email interface. In *Proceedings of the 2002 conference of the IBM Centre for Advanced Studies on Collaborative Research*, pages 136–145, Toronto, Canada, September 30 – October 3.

Hachey, Ben, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CONLL)*, pages 144–151, Stroudsburg, PA, USA, June 29–30.

Hair, Mario, Karen Renaud, and Judith Ramsay. 2007. The influence of self-esteem and locus of control on perceived email-related stress. *Computers in Human Behavior*, 23(6):2791–2803, November.

Halvey, Martin J. and Mark T. Keane. 2007. An assessment of tag presentation techniques. In *Proceedings of the 16th International Conference on the World Wide Web (WWW)*, pages 1313–1314, Banff, Alberta, Canada, May 8–12.

Hancher, Michael. 1979. The classification of cooperative illocutionary acts. *Language in Society*, 8(1):1–14, April.

Hassell, Lewis and Margaret Christensen. 1996. Indirect Speech Acts and Their Use in Three Channels of Communication. In *Proceedings of the First International Workshop on Communication Modeling - The Language/Action Perspective*, Tilburg, The Netherlands, July 1–2.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March.

Hiltz, Starr R. and Murray Turoff. 1985. Structuring computer-mediated communication systems to avoid information overload. *Communications of the ACM*, 28(7):680–689, July.

Hinkle, Steve, William B Stiles, and Laurie As Taylor. 1988. Verbal processes in a labour/management negotiation. *Journal of Language and Social Psychology*, 7(2):123–136, June.

Hoang, Quoc. 2012. Email statistics report 2012–2016 executive summary. Technical report, The Radicati Group Inc., April.

Hripcsak, George and Adam S Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, May / June.

Jackson, John Hughlings. 1874. On the duality of the brain. *Medical Press and Circular*, 17:19–22.

Jackson, Thomas W., Ray Dawson, and Darren Wilson. 2003. Understanding email interaction increases organizational productivity. *Communications of the ACM*, 46(8):80–84, August.

Jeong, Minwoo, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-Supervised Speech Act Recognition in Emails and Forums. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1250–1259, Singapore, 6–7 August.

Joachims, Thorsten. 1998. Text categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*, pages 137–142, Chemnitz, Germany, April 21–24.

Jones, William. 2008. *Keeping Found Things Found: The Study and Practice of Personal Information Management: The Study and Practice of Personal Information Management.* Morgan Kaufmann Publishers, San Francisco, CA, USA.

Joty, Shafiq, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1807–1813, Barcelona, Spain, July 16–22.

Joty, Shafiq, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2011. Supervised Topic Segmentation of Email Conversations. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 530–533, Barcelona, Spain, July 17–24.

Joty, Shafiq, Guiseppe Carenini, Gabriel Murray, and Raymond T Ng. 2010. Exploiting Conversation Structure in Unsupervised Topic Segmentation for Emails. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–398, Boston, MA, USA, 9–11 October.

Kaizer, Jasper and Anthony Hodge. 2005. AquaBrowser Library: search, discover, refine. *Library Hi Tech New*, 22(10):9–15.

Kerr, Bernard and Eric M Wilcox. 2004. Designing Remail: Reinventing the Email Client Through Innovation and Integration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 837–852, Vienna, Austria, April 24–29.

Khoo, Anthony, Yuval Marom, and David Albrecht. 2006. Experiments with Sentence Classification. In *Proceedings of the Australasian Language Technology Workshop (ALTW)*, pages 18–25, Sydney, Australia, November 30 – December 1.

Khosravi, Hamid and Yorick Wilks. 1999. Routing email automatically by purpose not topic. *Journal of Natural Language Engineering*, 5(3):237–250, September.

Klimt, Bryan and Yiming Yang. 2004. Introducing the Enron Corpus. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, USA, July 30–31.

Kline, Susan L., Cathy L. Hennen, and Kathleen M. Farrell. 1990. Cognitive complexity and verbal response mode use in discussion. *Communication Quarterly*, 38(4):350–360, Fall.

Kokkalis, Nicolas, Thomas Köhn, Carl Pfeiffer, Dima Chornyi, Michael S. Bernstein, and Scott R. Klemmer. 2013. EmailValet: Managing Email Overload through Private, Accountable Crowdsourcing. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, San Antonio, TX, USA, Feb 23–27. ACM.

Kraut, Robert E, Robert S Fish, Robert W Root, and Barbara L Chalfonte. 1992. Informal Communication in Organizations: Form, Function and Technology. In Ronald M Baecker, editor, *Readings in Groupware and Computer-Supported Cooperative Work: Assisting Human-Human Collaboration.* Morgan Kaufmann, pages 287–314.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology.* Sage Publications, Beverley Hills, CA, USA.

Labov, William and David Fanshel. 1977. *Therapeutic Discourse: Psychotheray as Conversation.* Academic Press, New York, NY, USA.

Lai, Eric. 1996. The Internet's Killer Application. In Wendy Woods, editor, *NewsBytes Executive Summary.* Island Telecommunications Corporation and Newsbytes News Network, December 9.

Lampert, Andrew, Robert Dale, and Cécile Paris. 2006. Classifying Speech Acts using Verbal Response Modes. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW)*, pages 34–41, Sydney, Australia, November 30 – December 1.

Lampert, Andrew, Robert Dale, and Cécile Paris. 2008a. Requests and Commitments in Email are More Complex Than You Think: Eight Reasons to be Cautious. In *Proceedings of Australasian Language Technology Workshop (ALTA)*, pages 55–63, Hobart, Australia, December 8–10.

Lampert, Andrew, Robert Dale, and Cécile Paris. 2008b. The Nature of Requests and Commitments in Email Messages. In *Proceedings of EMAIL-08: the AAAI Workshop on Enhanced Messaging*, pages 42–47, Chicago, IL, USA, July 13.

Lampert, Andrew, Robert Dale, and Cécile Paris. 2010. Detecting Emails Containing Requests for Action. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 984–992, Los Angeles, CA, USA, June 1–6.

Lampert, Andrew, Cécile Paris, and Robert Dale. 2007. Can Requests-for-Action and Commitments-to-Act be Reliably Identified in Email Messages? In *Proceedings of the 12th Australasian Document Computing Symposium (ADCS)*, pages 48–55, Melbourne, Australia, December 10.

Laver, John. 1981. Linguistic routines and politeness in greeting and parting. In Florian Coulmas, editor, *Conversational routine : explorations in standardized communication situations and prepatterned speech*. Mouton, pages 289–304.

Leech, Geoffrey N. 1983. *Principles of Pragmatics*. Longman Publishing Group, London, UK.

Leuski, Anton. 2004. Email is a stage: discovering people roles from email archives. In *Proceedings of the 27th Annual International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, pages 502–503, Sheffield, UK, July 25–29.

Mackay, Wendy E. 1988. More than just a communication system: Diversity in the use of electronic mail. In *Proceedings of the ACM conference on Computer-Supported Cooperative Work (CSCW)*, pages 344–353, Portland, Oregon, USA. MIT, ACM Press.

Macskassy, Sofus A., Aynur A. Dayanik, and Haym Hirsh. 1999. EmailValet: Learning User Preferences for Wireless Email. In *Proceedings of the Learning About Users and Machine Learning for Information Filtering Workshop at the International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, August 1.

Malinowski, Bronislaw. 1923. The problem of meaning in primitive languages. In Charles Kay Ogden and Ivor Armstrong Richards, editors, *The meaning of meaning: A study of the influence of language upon thought and the science of symbolism*. Routledge, London, pages 451–510.

McDonald, Ryan, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured Models for Fine-to-Coarse Sentiment Analysis. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 432–439, Prague, Czech Republic, June 24–29.

Meeuswesen, Ludwein, Cas Schaap, and Cees van der Staak. 1991. Verbal analysis of doctor-patient communication. *Social Science and Medicine*, 32(10):1143–50.

Mildinhall, John W. and Jan M. Noyes. 2008. Toward a stochastic speech act model of email behavior. In *Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, USA, 21–22 August. Conference on Email and Anti-Spam (CEAS) 2008.

Miller, Nancy L and William B. Stiles. 1986. Verbal familiarity in american presidential nomination acceptance speeches and inaugural addresses (1920-1981). *Social Psychology Quarterly*, 49(1):72–81, March.

Morris, Charles. 1946. *Signs, Language, and Behavior*. Prentice Hall, New York, NY, USA.

Murray, Denise E. 1991. *Conversation for Action: The Computer Terminal As Medium of Communication*. John Benjamins Publishing.

Neustaedter, Carman, A Brush, and Mark Smith. 2005. Beyond 'From' and 'Received': Exploring the Dynamics of Email Triage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1977–1980, Portland, OR, USA, April 2–7.

Pedersoli, Marco, Andrea Vedaldi, and Jordi González. 2011. A coarse-to-fine approach for fast deformable object detection. In *Procedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1353–1360, Colorado Spring, CO, USA, June 21–23.

Petrov, Slav. 2012. *Coarse-to-fine natural language processing*. Springer.

Petrov, Slav, Aria Haghighi, and Dan Klein. 2008. Coarse-to-fine syntactic machine translation using language projections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 108–116, Honolulu, Hawaii, USA, October 25–27.

Petrov, Slav and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 404–411, Rochester, NY, USA, April 22-27.

Piao, Scott. 2007. SentParBreaker Web Page. http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector, Accessed: 5/10/2007.

Piao, Scott S L, Andrew Wilson, and Tony McEnery. 2002. A Multilingual Corpus Toolkit. In *Proceedings of Fourth North American Symposium on Corpus Linguistics*, Indianapolis, IN, USA, November 1–3.

Pujara, Jay and Lise Getoor. 2010. Coarse-to-Fine, Cost-Sensitive Classification of E-Mail. In *Proceedings of the Workshop on Coarse-to-Fine Processing at the Twenty-Fourth Annual Conference on Neural Information Processing Systems (NIPS)*, Whistler, British Columbia, Canada, December 10.

Rak, Diana S and Linda M McMullen. 1987. Sex-role stereotyping in television commercials: A verbal response mode and content analysis. *Canadian Journal of Behavioural Science*, 19(1):25–39, January.

Reder, Stephen and Robert G. Schwab. 1990. The temporal structure of cooperative activity. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 303–316, Los Angeles, CA, USA, October 7–10.

Resnick, Peter W. 2008. Internet message format. RFC 5322, Internet Engineering Task Force (IETF), October.

Rittel, Horst and Werner Kunz. 1970. Issues as elements of information systems. Working paper 131, Institute for Urban and Regional Development, University of California at Berkeley, Berkeley, CA, USA, July.

Ritter, Alan, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 172–180, Los Angeles, CA, USA, June 1–6.

Robinson, Mike. 1990. Computer-Supported Cooperative Work and Informatics for Development. In *Proceedings of Informática*, Havana, Cuba, February.

Rohall, Steven. L and Dan Gruen. 2002. ReMail: A Reinvented Email Prototype. Demonstration at the ACM conference on Computer-Supported Cooperative Work (CSCW), November 16–20.

Sadock, Jerry M. and Arnold Zwicky. 1985. Speech act distinctions in syntax. In Timothy Shopen, editor, *Clause Structure*, volume 1 of *Language Typology and Syntactic Description*. Cambridge University Press, pages 155–196.

Sapp, Benjamin, Alexander Toshev, and Ben Taskar. 2010. Cascaded Models for Articulated Pose Estimation. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Proceedings of the 11th European conference on Computer vision: Part II (ECCV)*, volume 6312 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 406–420.

Scerri, Simon, Brian Davis, Siegfried Handschuh, and Manfred Hauswirth. 2009. Semanta - Semantic Email made easy. In *The Semantic Web: Research and Applications — Proceedings of the 6th European Semantic Web Conference (ESWC)*, volume 5554 of *Lecture Notes in Computer Science*, pages 36–50, Crete, Greece, May 31 – June 4. Springer Berlin Heidelberg.

Scerri, Simon, Gerhard Gossen, Brian Davis, and Siegfried Handschuh. 2010. Classifying Action Items for Semantic Email. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC)*, pages 3324–3330, Valletta, Malta, May 19-21.

Scerri, Simon, Siegfried Handschuh, and Brian Davis. 2008. The path towards semantic email: Summary and outlook. In *Proceedings of EMAIL-08: the AAAI Workshop on Enhanced Messaging*, pages 64–70, Chicago, IL, USA, July 13.

Scerri, Simon, Myriam Mencke, Brian David, and Siegfried Handschuh. 2008. Evaluating the Ontology powering sMail — a Conceptual Framework for Semantic Email. In *Proceedings of the Sixth Conference on International Language Resources and Evaluation (LREC)*, pages 2640–2646, Marrakech, Morocco, May 28–30.

Schlegloff, Emanuel A. 1972. Notes on a conversational practice: Formulating place. In David Sudnow, editor, *Studies in Social Interaction*. Free Press.

Schlegloff, Emanuel A and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327, January.

Schölkopf, Bernhard and Alexander J Smola. 2002. *Learning With Kernels — Support Vector Machines, Regularization, Optimization and Beyond*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK.

Searle, John R. 1975. Indirect speech acts. In Peter Cole and Jerry L Morgan, editors, *Syntax and Semantics Volume 3: Speech Acts*. Academic Press, New York, NY, USA, pages 59–82.

Searle, John R. 1976. A classification of illocutionary acts. *Language in Society*, 5(1):1–23, April.

Searle, John R. 1979. *Expression and Meaning*. Cambridge University Press, Cambridge, UK.

Searle, John R. 1989. How performatives work. *Linguistics and Philosophy*, 12(5):535–558, October.

Sinclair, James and Michael Cardew-Hall. 2008. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, February.

Sinclair, John and Richard Malcolm Coulthard. 1975. *Towards and Analysis of Discourse - The English used by Teachers and Pupils*. Oxford University Press, London, UK.

Sproull, Lee and Sara Kiesler. 1991. *Connections: New ways of working in the networked organization*. MIT Press, Cambridge, MA, USA.

Steinfeld, Aaron, S. Rachael Bennett, Kyle Cunningham, Matt Lahut, Pablo-Alejandro Quinones, Django Wexler, Dan Siewiorek, Jordan Hayes, Paul Cohen, Julie Fitzgerald, Othar Hansson, Mike Pool, and Mark Drummond. 2007. Evaluation of an integrated multi-task machine learning system with humans in the loop. In *Proceedings of the 2007 Workshop on Performance Metrics for Intelligent Systems (PerMIS)*, pages 168–174, Gaithersburg, Maryland, USA, August 28–30.

Stiles, William B. 1992. *Describing Talk: a taxonomy of verbal response modes*. SAGE Series in Interpersonal Communication. SAGE Publications.

Stiles, William B and M Lauren White. 1981. Parent-child interaction in the laboratory: Effects of role, task, and child behavior pathology on verbal response mode use. *Journal of Abnormal Child Psychology*, 9(2):229–241, June.

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–371, September.

Sumner, Mary. 1988. The impact of electronic mail on managerial and organizational communications. In *Proceedings of the Conference on Office information Systems (COCS)*, pages 96–109, Palo Alto, CA, USA, March 23–25.

Thomas, Gail Fann, Cynthia L. King, Brian Baroni, Linda Cook, Marian Keitelman, Steve Miller, and Adelia Wardle. 2006. Reconceptualizing e-mail overload. *Journal of Business and Technical Communication*, 20(3):252–287, July.

Thompson, Sandra A and William C Mann. 1987. Rhetorical structure theory: A framework for the analysis of texts. *IPRA Papers in Pragmatics 1*, 1(1):79–105.

Tomlinson, Ray. 2006. The first network email. BBN. http://openmap.bbn.com/~tomlinso/ray/firstemailframe.html. Accessed: 8th March 2013.

Tong, Simon and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, November.

Tsai, Richard Tzong-Han, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7(1):92, February.

Tsui, Amy B M. 1989. Beyond the adjacency pair. *Language in Soci*, 18(4):545–564, December.

Tsui, Amy B M. 1994. *English Conversation*. Oxford University Press, London, UK.

Ulrich, Jan, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2009. Regression-based summarization of email conversations. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 334–337, San Jose, CA, USA, May 17–20.

Ulrich, Jan, Gabriel Murray, and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *Proceedings of EMAIL-08: the AAAI Workshop on Enhanced Messaging*, pages 77–82, Chicago, IL, USA, July 13.

Van Lancker, Diana. 1987. Non-propositional speech:neurolinguistic studies. In Andrew W Ellis, editor, *Progress in the Psychology of Language*, volume 3. Lawrence Erlbaum, Hillsdale, NJ, USA, pages 49–118.

Vanderveken, Daniel. 1990. *Meaning and speech acts: principles of language uses*, volume 1. Cambridge University Press, New York, NY, USA.

Venolia, Gina Danielle, Laura Dabbish, JJ Cadiz, and Anoop Gupta. 2001. Supporting email workflow. Technical Report MSR-TR-2001-88, Microsoft Research, Redmond, WA, USA, Revised December (Originally September). Collaboration and Multimedia Group.

Verschueren, Jef. 1983. Speech act classification: A study in the lexical analysis of english speech activity verbs. *Language*, 59(1):166–175, March.

Veselinova, Ljuba and Helen Dry. 1995. Queries on the linguist list: Acquisition of a subregister. Paper presented at the Georgetown University Round Table, Washington, D.C.

Vine, Bernadette. 2004. *Getting Things Done at Work*, volume 124 of *Pragmatics & Beyond New Series*. John Benjamins Publishing Company.

Wahlster, Wolfgang, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Artificial Intelligence. Springer, New York, NY, USA.

Waldvogel, Joan. 2002. Some features of workplace emails. *New Zealand English Journal*, 16:42–52.

Whittaker, Steve. 2005. Supporting collaborative task management in e-mail. *Human-Computer Interaction*, 20(1 and 2):49–88, June.

Whittaker, Steve, Victoria Bellotti, and Jacek Gwizdka. 2006. Email in personal information management. *Communications of the ACM*, 49(1):68–73, January.

Whittaker, Steve, Victoria Bellotti, and Paul Moody. 2005. Introduction to this special issue on revisiting and reinventing email. *Human-Computer Interaction*, 20(1 and 2):1–9, June.

Whittaker, Steve, David Frohlich, and Owen Daly-Jones. 1994. Informal workplace communication: what is it like and how might we support it? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 131–137, Boston, Massachusetts, USA, April 24–28.

Whittaker, Steve and H Schwarz. 1995. Back to the future: pen and paper technology supports complex group coordination. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 495–502, Denver, CO, USA, May 7–11.

Whittaker, Steve and Candace Sidner. 1996. Email Overload: exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 276–283, Vancouver, British Columbia, Canada, April 13–18.

Winograd, Terry. 1987. A language/action perspective on the design of cooperative work. *Human-Computer Interaction*, 3(1):3–30, March.

Winograd, Terry and Fernando Flores. 1986. *Understanding Computers and Cognition.* Ablex Publishing Corporation, Norwood, New Jersey, USA, 1st edition.

Witten, Ian and Eiba Frank. 2005. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, CA, USA, 2nd edition.

Yee, Ka-Ping. 2002. Zest: Discussion mapping for mailing lists. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 123–126, New Orleans, LA, USA, November 16–20.

Zhang, Haijun and Tommy W. S. Chow. 2011. A coarse-to-fine framework to efficiently thwart plagiarism. *Pattern Recognition*, 44(2):471–487, February.