Document Restructuring to Promote Semantic Coherence in Topics

By

Fiona Martin

A thesis submitted to Macquarie University
for the degree of
Masters of Research
Department of Computing
February 2015

MACQUARIE
UNIVERSITY
FACULTY OF SCIENCE

Typeset in LaTeX 2_ε.

Declaration

I certify that the work in this thesis entitled DOCUMENT RESTRUCTURING TO PROMOTE SEMANTIC COHERENCE IN TOPICS has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree to any other university or institution other than Macquarie University. I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Fiona Martin

Acknowledgements

I would like to thank the authors of *Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality* (2014), Jey Han Lau, David Newman and Timothy Baldwin, for making available the software tools for evaluating the observed coherence of topics and their automated word intrusion detection test.

I would like to thank Zhendong (Tony) Zhao for making available his Information Retrieval evaluation software for this study.

Finally, I would like to thank my supervisor, Professor Mark Johnson, for his guidance and encouragement during the preparation of this thesis.

Abstract

This thesis examined whether simple preprocessing of documents such as lemmatising text, or removing or weighting certain parts of speech, could generate better quality topics, faster, using Latent Dirichlet Allocation (LDA) topic modelling. Past work has generally attempted to improve topic modelling performance by making changes to the topic modelling algorithm itself. This study examines the simpler option of transforming the input documents to the LDA algorithm. Topic quality was assessed on a range of measures that examined both topic interpretability, and how well the topics represented the source documents. The results indicate that topic quality was improved, and the time to generate the topics was less, if the input documents were reduced to only nouns, or nouns and adjectives, when the numbers of topics to be generated was 200 or 500 topics. This study also found that even when the number of topics to generate was not large, input documents could be reduced to select parts of speech to speed the generation of topics, with no loss of topic quality. The implications of these results are that very large data sets may benefit from being lemmatised and reduced to simply the nouns prior to topic modelling.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Context of this Study

The increasing volume of digitised text has created new opportunities to both confirm past assumptions and for new knowledge discovery. To support these new lines of research, there is a need for automated methods to analyse and organise large document collections. One technique that has been of interest to fields as diverse as the digital humanities and biomedical research, is topic modelling. The goal of this study is to consider ways in which topic modelling can be enhanced to allow valid, reliable and meaningful insights to be learned from large, digitised corpora for applied fields of study such as the digital humanities.

Topic modelling is a form of analysis that seeks to identify latent themes within digitised collections. Such latent themes are referred to as topics. For collections of text documents, topics are in the form of word lists such as {*leaf*, *plant*, *soil*, *feed*, *water*}. The allocation of words to topics, and linking topics to documents, is based on probability theory, implemented through statistical algorithms. One of the most prominent of these algorithms is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). An attractive feature of algorithms such as LDA for text documents, is that words and documents are associated to all the topics in varying proportions, which allows documents to form overlapping clusters. This recognises that documents can contain multiple themes, and can contain themes to varying extents, as each word in each document links to a topic. While it is common for topics to be expressed as a list of the top ten or so most frequent words associated with the topic, the topics themselves are actually distributions across the entire vocabulary in the corpus.

Topic modelling has commonly been used as a pre-processing step, as a means of arranging documents to aid downstream information retrieval processes (Blei, 2012). However, as fields of study such as the digital humanities seek new ways to analyse digitised texts, there has been interest in whether the topics generated by topic modelling algorithms can be directly read to identify themes and patterns in such texts. Novel applications of topic modelling that directly use the generated topics to gain insight into the source texts include mining historical newspaper articles regarding the American Civil war (Nelson, 2010), measuring expressed agendas in U.S. Senate Press Releases (Grimmer, 2010), and to use geodata from digitised 19th century ships logs to examine shipping

routes (Schmidt, 2012). In each of these studies, the researcher has applied freely available topic modelling tools to their respective datasets, adjusting the data as required to meet the requirements of the tool, an approach common in the digital humanities domain (Brett, 2012) [1]. The more direct application of topic modelling for data analysis brings with it different qualifications of what makes a good topic.

As an upstream process to information retrieval tasks, topics provide a means to access large, unstructured data sets. There is no requirement that the topics are interpretable to human eyes, or that the most frequent words in the topic are actually the most important to identifying the underlying theme to the topic. The overriding requirement is that the topics facilitate information retrieval. As the topics typically cover many thousands of words, and documents belong to multiple topics with varying probabilities, then if the topic is actually a conjunction of two (or more) themes, then this is not particularly disruptive for information retrieval. Such conjunctions can, however, produced incomprehensible word lists, or lists that can be misleading when only the top ten or so words in a topic are viewed (Mimno et al., 2011).

The true theme behind a topic may require consideration of more than the top ten most frequent topic words. A neat example of this is shown in Schmidt (2012)'s shipping log study, where plotting only ten co-ordinates from each topic of geodata co-ordinates gives a poor representation of the shipping routes in each topic. Only by plotting far more points (in this study, 25), was the route more fully represented. This would be equivalent to trying to derive meaning from the top ten words of a topic generated from a newspaper corpus, when a review of the top twenty or thirty words would suggest a completely different meaning. The challenge for the human reviewer is that the more words viewed for a topic, the more difficult it is for the reviewer to collate all the words into a meaningful theme.

To be useful to gain insights into a corpus, the generated topics need to be meaningful, valid and reliable indicators of themes in the corpus. If the topic is to be directly analysed, rather than as an input into an information retrieval process, then to be meaningful the words in the topic need to combine to suggest an interpretable theme. To this end, recent attention has focussed on the semantic coherence of topics, where semantically coherent topics have a set of most frequently occurring words that combine to convey an interpretable theme. For example {*water plant tree garden flower fruit valley drought*} are semantically coherent, whereas {*art museum house room style work fashion water*} seem to combine multiple themes. Additionally, topics can be coherent without conveying a useful meaning. For example, a topic of {*told asked called time phone calls call heard*} may be coherent, but the lack of entities or concepts makes such a topic not particularly useful to applied fields such as the digital humanities or bio-medical text mining.

In seeking ways to evaluate the semantic coherence of topics, recent work has focused on how topic coherence can be assessed by comparing generated topics to word co-occurrence in the source texts (for example, in the work by Lau et al. (2014) to be

[1] Other examples of topic modelling in the digital humanities field can be found at http://programminghistorian.org/lessons/topic-modeling-and-mallet (last reviewed 20th January, 2015)

discussed later in this study). The extension to simply measuring topic coherence is to proactively make topic modelling generate more semantically coherent topics, as in Mimno et al. (2011). Traditionally, computer scientists approach improving topic modelling by proposing alternative algorithms (as in Mimno et al. (2011)). However, there are two parts to topic modelling, the algorithm and the data. In the applied use of topic modelling, adjusting the data is more readily accessible to most researchers than reconfiguring the modelling algorithm. The research question of interest in this study was whether documents could be restructured to produce more semantically coherent topics. For example, if documents were restructured to place greater weight on the nouns (for example, museum, water, plant, style, Australia) compared to other parts of speech, would the topics be more meaningful to a reviewer seeking to understand the main themes in a corpus.

1.2 Objective of this Study

The aim of this study was to determine if more semantically coherent topics could be generated by restructuring the documents to place greater emphasis on the elements in the text that detail the key entities and concepts in each document, that is the semantic content of the documents, while using a standard, readily available topic modelling tool to generate the topics.

In this study, semantic content is assumed to be conveyed in the nouns and named entities in the source text. In particular, this study will examine a diverse, multi-themed corpus of newspaper articles from the 1991 San Jose Mercury News, published in the Tipster corpus. In such a corpus, generating coherent topics is challenging because of the wide range of entities and concepts referenced, in the broad collection of articles. The general hypothesis considered was whether restructuring such news articles to place greater weight on the nouns[2] or named entities produced more semantically coherent topics than generating topic models from the original, unaltered, articles. A range of forms of restructuring around nouns and named entities were trialled, and in each trial the hypothesis tested was that such restructuring improved topic coherence over generating topic models from the articles in their original form. Semantic coherence was assessed using measures made available by the research of Lau et al. (2014).

The significance of this study is that in manipulating the data rather than adjusting the algorithm, this study uses an approach more readily accessible to the general research community. The tools to be used to generate topic models, and to identify elements such as nouns and named entities, were software tools that were readily available to the broader research community.

1.3 Overview of the Study

The following chapters provide details of the forms of document restructuring that were evaluated in an attempt to generate more semantically coherent topics, and what the

[2]Variants of nouns with adjectives, or nouns with adjectives and verbs were also considered

outcomes of those trials were. Before detailing the experiments, the Background chapter provides a brief overview of the forms of restructuring to be used in this study, including a brief overview of lemmatisation, part of speech tagging, named entity tagging, and a discussion of the structures sometimes found in newspaper articles. The Literature Review chapter describes how part of speech tagging, entities, and document structure have been combined with topic modelling in a selection of relevant past studies. The Method chapter details the models trialled in this study, how documents were restructured for each trial, and how the various trials were compared. The Results chapter examines the complexities found in the data set that was topic modelled, and presents the results for each quantitive evaluation undertaken as well as a brief qualitative review of the topics produced in select trials. The Discussion chapter reviews the usefulness of each form of restructuring trialled, and the Conclusion chapter summarises the findings of this study, and examines future extensions to improving the semantic coherence of topics when the goal is to produce topics that are meaningful for direct review.

Chapter 2

Background

2.1 Introduction

In answering the question of whether documents can be restructured to produce more semantically coherent topics, first the ways that documents could be restructured need to be considered. This chapter will provide an overview of lemmatisation, part of speech tagging, named entity recognition and will discuss the structure of newspaper articles. Finally, this chapter will provide an overview of the type of topic modelling used in this study, Latent Dirichlet Allocation (LDA).

2.2 Lemmatisation

In text mining it is often desirable to reduce words back to a stem or lemma, removing tense or declension. For example, in English language texts, the Porter Stemmer (Porter, 1980, 1997) is often used to remove prefixes, suffixes and inflections from words. However, the resultant stems are more difficult for the average human reader to scan than the original words. Lemmatisation is the process of mapping a word form to a lemma, such as reducing *found* to *find* (where stemming would have left this as *found*). The advantage of lemmatisation is the result can be easier to read than the results from stemming. However, neither stemming nor general lemmatisation are useful to reduce gamma-amino-butyric acid and GABA to a common form. That is, standard lemmatisation routines do not cater for specialist vocabularies.

2.3 Part of Speech Tagging

Part-of-speech (POS) tagging divides words into word classes (also referred to as morphological classes or lexical tags), and assigns each word a type tag. For example, POS tagging will separate nouns from verbs, and within nouns, may separate common nouns from proper nouns, and singular from plural nouns. Several well known tag sets have been defined, and for texts written in English include the 87-tag Brown corpus (Francis & Kucera, 1979) and the 45-tag Penn Treebank (Marcus et al., 1993).

Of interest in this study is the way POS tags can be used to separate function words from content words. While the function words (prepositions, determiners, pronouns,

conjunctions and so on) are important for grammatically correct sentences, it is the content words (the nouns, verbs, adjectives and adverbs) that convey the who, what, when and how, that is of most interest to text mining (outside perhaps the field of linguistics). It is this semantic aspect of content words that makes them most useful for identifying topics and themes associated with entities and concepts in documents. Many topic models implicitly focus on content words by excluding the function words through the use of stop lists. Stop lists are a predefined list of high-frequency words judged to be uninformative were they to be included in a topic, words such as 'and', 'the', or 'of'. Stop lists do not considered word-sense, and so cannot distinguish the multiple uses of 'will' to capture acts of *will*, and a legal *will*, while ignoring the potentially highly frequent *will* as a verb. Consideration of POS could, however, allow a distinction at least between the noun and the verb.

There are many open source software tools available to perform POS tagging. An example is the Stanford POS Tagger (Toutanova et al., 2003). The Stanford POS Tagger is a maximum-entropy (CMM) part-of-speech (POS) tagger for English, Arabic, Chinese, French, and German, and is a Java implementation of log-linear part-of-speech tagging. The English tagger uses the Penn Treebank POS tags. While POS taggers like the Stanford POS Tagger typically average over 97% token accuracy (Manning, 2011) with the inclusion of non-ambiguous tokens like punctuation marks, this accuracy can be reduced down to approximately 57% (Manning, 2011) when looking at whole sentences, rather than tokens.

2.4 Named Entity Recognition

'Named entity' is an information retrieval term that refers to proper nouns for entities. Such proper nouns tend to be categorised as person, location, organisation, times, amounts and so on. Such categorisations form meta-data to be associated with the entity. The rules for classifying such proper nouns, and how many categories are used, is typically application specific (Jurafsky & Martin, 2009). For example, the Stanford Named Entity Recognizer (Finkel et al., 2005) is a Java implementation of a Named Entity Recogniser, that provides the option of choosing 3, 4 or 7 categories (3 class: Location, Person, Organisation; 4 class: Location, Person, Organisation, Misc; 7 class: Time, Location, Organisation, Person, Money, Percent, Date) trained on English newswire data, with the option for constructing additional feature extractors for other types of word sequences. The Stanford Named Entity Recognizer (Stanford NER) is a linear chain Conditional Random Field (CRF) sequence model implementation. As such, it is discriminative rather than generative, makes no assumptions regarding feature independence, and uses global rather than local normalisation (Finkel, 2007). A limitation of such an approach is that it may be slower to process data compared to alternatives (Finkel et al., 2005) such as Hidden Markov Model (HMM) implementations, and Maximum Entropy Markov Models (MEMMs).

Regardless of the number of categories used by a particular NER algorithm, identification of named entities can be challenging when different terms or styles are used

to reference the same underlying entity. One of the main challenges for named entity recognition is co-reference resolution, where the same entity is referred to using different noun phrases, such as mixing *J. Smith*, *Jane Smith*, *she* and *Ms Smith* for the same person. Another challenge is to resolve instances where the same name refers to different entities, for example, where there are multiple towns and suburbs named *Richmond*.

2.5 Document Structure

The sequencing of sentences within documents, and the division of documents into sections, can also offer useful cues to the content of a document. Patterns across the sentences within a text are sometimes referred to as *discourse structures* (Webber et al., 2012), and the flow of the discourse can follow a very genre-specific conventional structure. For example, many fictional stories have a hero, a challenge, and a successful resolution, as an example of discourse structures by eventualities. News reports can have more function-based discourse structures (Webber et al., 2012). For example, news articles have an inverted pyramid structure (Webber et al., 2012), where the lead paragraph introduces the key actors, events, motivations and locations in the story, and as the article proceeds, each of these items is expanded upon in order of importance, with the least important information at the end of the story.

Understanding conventionalised discourse structures can aid extracting information from a document. The experienced reader scans the appropriate section before deciding whether to read on or not. If the goal is to capture the most salient aspects of a news story, it may be that the lead paragraph lists the key people, places, events and motivations. This is an aspect of document structure that will be investigated in this study.

2.6 Topic Modelling

The key premise behind topic modelling is that documents are assumed to be a mix of latent topics, and the high frequency words in each topic tend to co-occur in documents containing that latent topic. For example, the word set { *tree*, *plant*, *soil*, *feed*, *water* } may occur to differing degrees in a variety of stories in a corpus of newspaper articles, and could subjectively be titled *gardening*. Articles containing such words are considered linked to this latent topic of gardening, to a greater or lesser degree. For example, an article about eating at a restaurant may be linked to the gardening topic if it contains the word 'feed', but it will not be as strongly associated to this topic as an article that contains 'plant', 'feed' and 'water'. While we may think of topics as comprising entities or concepts such as gardening or governance, the topic modelling algorithm just produces an association of words. The algorithm is just as likely to produce the gardening topic as it is to produce {*told asked called time phone calls call heard*}. In fact, this later topic may be more likely, since such words may occur more frequently in particular documents and data sets than a particular form of reference to an entity.

A more formal definition of topics provided by Blei (2012), is that a topic is a multinomial distribution from a fixed vocabulary. As an unsupervised machine learning

algorithm, topic modelling provides a means of analysing large digital corpora in ways not possible by manual review. Topic modelling can be used to classify or categorise large volumes of documents, for exploring themes and writing styles across collections, and across time, and can also provide a means for retrieving information from documents using a more abstract or generic method than is possible using specific keyword searches.

One of the most predominant, and simple (Blei, 2012), topic models is Latent Dirichlet Allocation (LDA). The LDA algorithm develops topics based on statistical measures for posterior probabilities, rather than to fit a particular end-user search criteria. The word to topic probabilities and topic to document probabilities are calculated using a Dirichlet distribution as a prior (see Blei et al. (2003) for a full explanation of the algorithm, and Blei (2012) for a simplified explanation). Blei et al. (2003) proposes that such an approach allows LDA to perform better in prediction of topics for held-out documents than it's main competitor Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999), which is more prone to over fitting to the training data.

The LDA model operates under several assumptions (Blei et al., 2003), including that the number of topics to generate is specified in advance, and that word order and document order can be ignored. Variants of LDA have been developed when violation of these assumptions is required, including non-parametric topic modelling to relax the assumption of knowing the number of topics in advance (for example, see Teh et al. (2006)), and Dynamic Topic Models (Blei & Lafferty, 2006) when tracking themes over time, and therefore document sequence, is of interest. Of interest to this current study is word order, as used in POS taggers to identify nouns from verbs, and in word collocations that identify 'South Australia' as a different entity from 'Australia'. That is, rather than treat documents as *bags of words* (Salton & McGill, 1983)[1] as LDA does, this study seeks to use word order to focus on those elements that provide semantic content. There are several alternative topic modelling algorithms that similarly violate the bag of words assumption, and these will be reviewed in the next chapter. However, these algorithms, like Blei & Lafferty (2006)'s Dynamic Topic Model, modify the topic modelling algorithm to circumvent limiting assumptions of the LDA approach.

The LDA algorithm provides a flat representation of topics. Other topic modelling algorithms that allow for a more sophisticated, interrelated view of topics include the hierarchical Dirichlet process, proposed by Teh et al. (2006) for data with a predefined hierarchical structure, the Correlated Topic Model, proposed by Blei & Lafferty (2007) that allows limited correlations between pairs of topics, the hierarchical LDA (Blei et al., 2004) that allows hierarchical structures between topics, but restricts the document sampling to individual branches of the hierarchical structure (Li & McCallum, 2006), and the Pachinko Allocation Model (PAM) algorithm that considers relationships between topics, in addition to the word to topic relations (Li & McCallum, 2006). While such algorithms may offer advantages to dealing with a newspaper corpus containing multiple, interrelated themes, LDA is one of the most widely used algorithms (Blei, 2012), and it will be the topic generation method used in this study.

[1] *bag of words*: Where a document or text is seen as a collection of words where word order is unimportant

2.7 Conclusion

This chapter has provided an overview of lemmatisation, POS tagging, NER, LDA, and briefly discussed some standards for structuring news articles. These are techniques for dealing with text data that will be applied or considered in this study, where the goal is to determine if documents can be restructured using such approaches to produce more semantically coherent, valid and reliable topics.

Chapter 3

Literature Review

3.1 Introduction

As the types of restructuring to be considered in this study are centred around the use of select parts of speech, of named entities, and of document structure, it is appropriate to consider how these elements have been combined with topic modelling in the past, and how such work informs this current study. This chapter will consider studies that have sought to integrate parts of speech with topic modelling, and the recognition of entities with topic modelling. Additionally, studies that have considered a role for document structure when topic modelling will be considered.

3.2 Part of Speech and Topic Modelling

Latent Dirichlet Allocation (LDA) is a bag of words approach to topic modelling, but several alternative algorithms have been proposed that seek to treat words differently based on syntactic class. For example, the combined Hidden Markov - LDA model (HMMLDA) (Griffiths et al., 2005), the Part-Of-Speech LDA model (POSLDA) (Darling et al., 2012), and the nested HMMLDA model (nHMMLDA) (Jiang, 2009) all integrated part of speech (POS) into the topic modelling algorithm. All seek to recognise that semantic content is provided by the open class words (nouns, verbs, adjectives, adverbs), not the closed class, function words. In each of these three approaches, content words are identified by using a Hidden Markov Model (HMM) to identify the POS of each word.

An alternative approach proposed by Boyd-Graber & Blei (2009) sought to use pre-defined sentence dependency parse trees to identify word use, and then groups tokens based on their patterns of use. Boyd-Graber & Blei (2009) termed their model a Syntactic Topic Model (STM), and the motivation behind this model was a desire to predict the types of words that may fit with particular sentence patterns (Boyd-Graber & Blei, 2009). The topics produced tended to group words into topics such as {*runs*, *falls*, *walks*, *sits*, *climbs*} (Boyd-Graber & Blei, 2009), seeming to implicitly group like parts of speech, though not as definitively as the HMM approaches.

These various algorithms attempted to produce less noisy, more semantically coherent topics by grouping content words in topics, and isolating function words from content words. In the HMMLDA model, function words are drawn into a separate, corpus level

topic (Griffiths et al., 2005), and the content words can then be used to generate a separate set of content based, less noisy, topics. The POSLDA approach allowed topics to be created by POS, grouping nouns into one set of topics, verbs into another set, and adjectives into a third set, allowing comparisons across domains of typical word use. The nHMMLDA model similarly split syntactic and semantic aspects of text, but assumed that words within sentence boundaries are all related to a single topic, whereas POSLDA allowed any individual word to be associated with any topic.

In each of the above alternative topic modelling algorithms, the respective authors suggest topic quality is enhanced by focussing on content words. The nHMMLDA, POSLDA and STM were all assessed using a measure of perplexity on a held-out data set (Darling et al., 2012; Griffiths et al., 2005; Jiang, 2009), whereas the HMMLDA was assessed using a method based on the harmonic mean of likelihoods (Boyd-Graber & Blei, 2009). It is useful here to consider the work of Chang et al. (2009) that these statistical measures do not necessarily indicate that the topics are interpretable to human reviewers. An open question is whether the usefulness of focussing on content words also can be demonstrated with different evaluation techniques such as measures of the semantic coherence of topics.

The gold standard for evaluating topic coherence is human review (Chang et al., 2009). However, even with a modest number of topics, it can be time consuming and attention taxing to review lists of the top ten words for each topic, which has motivated several studies into automated evaluation of topic coherence. To automate the evaluation of topic coherence requires some criteria for judging coherence. Recent studies such as Newman et al. (2010), Mimno et al. (2011), Aletras & Stevenson (2013) and Lau et al. (2014) have examined the role of word co-occurence in the corpora of interest compared to a separate, but relevant, reference corpus. The premise of such studies is that if it is assumed that words that are related tend to co-occur in corpora (Harris (1954) as cited in Aletras & Stevenson (2013)), and if the most frequent words in a topic co-occur in a separate but relevant reference corpus, then this may indicate the topic is coherent, and interpretable. Each of these studies found correlation between their individual measures of word co-occurrence and human evaluations of topic coherence, though the forms of the coherence measures differed by study. For example, Newman et al. (2010) used a point-wise mutual information measure (PMI), Mimno et al. (2011) used a log conditional probability measure, while Aletras & Stevenson (2013), and Lau et al. (2014) also used a normalised point-wise mutual information (NPMI) measure (developed in Bouma (2009)). The NPMI measure was found by Lau et al. (2014) and Aletras & Stevenson (2013) to be less susceptible to bias for lower frequency words, and yielded more consistent results on topic level evaluations than the other measures trialled in each respective study. Furthermore, Mimno et al. (2011) found that since topic modelling did not use word co-occurrence, there was not a requirement to use an external or held out reference corpus, that co-occurrence could be assessed using the topic modelled data set.

3.3 Entities and Topic Modelling

Combining entities with topic models has taken several different paths. Firstly, the problem of resolving and disambiguating multiple forms of reference to the same underlying entity (e.g. resolving 'Apple Inc' with 'Apple' the company, and separating references to 'apple' the fruit) has been the subject of several interesting studies that use the frequency of words and their associations in topics for disambiguation (for example, see Bhattacharya & Getoor (2006), Han & Sun (2012) and Shu et al. (2009)). Secondly, Newman et al. (2006) explored indirect entity relationships through topic associations. Thirdly, building topics for specific, nominated entities has been explored in Kim et al. (2012) as a means of understanding the themes associated with particular entities. The Kim et al. (2012) study used semi-supervised rather than unsupervised learning, selecting articles using a keyword search relevant to the multiple entities of interest, disambiguating multiple forms of reference to the entities, then generating topics models on this preprocessed data set.

Each study above developed new algorithms to integrate entity handling with topic models. Applied fields of study such as Yang et al. (2011)'s modelling of historic newspapers also sought to identify named entities in the source text, but rather than change the algorithm, the Yang et al. (2011) study ran named entity recognition as a separate pre-processing step. In different ways, each of the approaches mentioned here seeks to manage the complexities in language, where entities (and concepts) can be referenced using multiple, interchangeable terms, and that sometimes those terms are ambiguous (e.g. Apple as a company and as a fruit).

3.4 Document Structures and Topic Modelling

Document structures have been analysed and integrated with topic models in studies such as Zhu et al. (2006), Du et al. (2012), Wang et al. (2011) and Chen et al. (2009). Each study, in its own way, indicates that being discourse "structure aware"(p.371), as Chen et al. (2009) terms it, can produce topics that better reflect the sections from which the topic was generated. As indicated by the work of Zhu et al. (2006), each distinct section can have its own themes, and as indicated in the work of Du et al. (2012), there can be a sequential progression of topics through a text. Both sets of findings were based on longer works (research papers and novels). Of interest for this study was whether discourse structure patterns can be applied in newspaper articles, and in particular, whether the lead paragraph summarises the key entities and concepts that represent the themes of a newspaper article, and whether weighting these elements in the lead paragraph may produce more meaningful document-topic relationships.

3.5 Conclusion

In summary, past studies have identified that violating the bag of words assumption to account for word use (i.e. content or function), word collocations, and word location

by document section, can add value in topic modelling. This suggests not all words necessarily have equal value when generating useful topics. Each of the models discussed in this chapter customised the topic modelling algorithm (requiring programming skills to implement such changes), and topic quality has typically been assessed using perplexity based measures, rather than the semantic coherence approaches discussed in recent studies such as Newman et al. (2010), Mimno et al. (2011), and Lau et al. (2014). Therefore, one open question is whether restructuring the source documents rather than customising the topic modelling algorithm, and assessing topic coherence rather than using a perplexity based measure, can find similar advantages to emphasising content words, particular word collocations for named entities, and consideration of document structure.

Chapter 4

Method

4.1 Introduction

The research question of interest in this study was whether restructuring documents to highlight the aspects that convey semantic content would generate more semantically coherent topics. The restructuring of interest was based on content words, named entities, and the distinction between the first versus subsequent paragraphs in the text. Furthermore, the content of interest in this study related to entities and concepts, rather than actions, with more focus on nouns than on verbs alone. The aim was to determine if a particular form of document restructuring would produce topics that were assessed to be more semantically coherent on a quantitative measure of coherence.

Various forms of document restructuring were tested, with separate data sets created for each. An overview of this restructuring is shown in Figure 4.1, and will be explained in detail in the subsequent sections of this chapter[1]. Topics were generated for each of the individual data sets, and the quality of the generated topics was assessed using several measures. The semantic coherence of individual topics was of most interest, and this was evaluated using two quantitative measures from a study by Lau et al. (2014) that assessed whether the co-occurrence of particular words in a topic reflected the actual co-occurrence of such words in the source data. Having topics that consist of a set of words commonly found co-occurring may be a sign of an interpretable theme in that topic. While semantically coherent topics are important for being able to interpret themes from a topic, it is also important that the topics are reflective of the source data. In this regards, this study sought to examine document to topic associations using both an information retrieval measure, and by examining how well individual topics were reflected in individual documents, as rated by the topic modelling tool. Qualitative aspects of the generated topics were also examined, however, it was difficult to effectively perform substantial qualitative reviews when the source data set was sizeable and the number of topics was large.

This chapter will provide details on the data set that was the basis for all trials, the means of restructuring and topic modelling the data, and the various evaluations employed. Note, this chapter will use the word 'trial' to refer to each restructured data set and the topic models generated over that data set.

[1]Note that none of the trials combined both part-of-speech tagging and named-entity identification

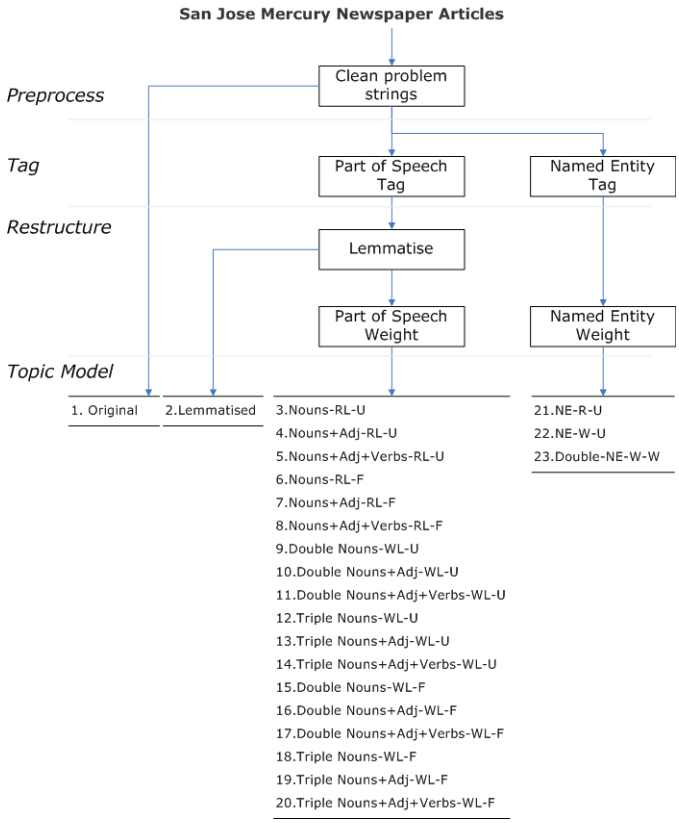


FIGURE 4.1: Overview of the process to generate the trial data sets
Note: Detailed explanations of the 22 datasets are provided in Table 4.1

4.2 Data

Topic models were generated from a set of San Jose Mercury News (SJMN) articles, available in the Tipster corpus[2] (third disk). The articles in this corpus are in a manually annotated, standard SGML format. The SGML tags of interest to this study were the <HEADLINE>, <LEADPARA>and <TEXT>, where the lead paragraph of the article has been separated from the main text of the article. The Tipster corpus did not identify individual sentences, parts of speech or paragraph breaks other than separating the lead paragraph. Due to memory restrictions in both the part of speech tagger and the named entity tagger, the SJMN articles were grouped into 38 sets of no more than 2501 articles per set, split using a Python script and the ElementTree XML API. ElementTree required the replacement of '&plus' with '+', and '&equals' with '=' in the original SJMN text.

[2]https://catalog.ldc.upenn.edu/LDC93T3A

4.3 Part-of-Speech Tagging

Part of speech (POS) tagging was performed using Stanford Log-linear Part-of-Speech tagger (StanfordPOS) (Toutanova et al., 2003), v3.3.1 (2014-01-04), a maximum-entropy (CMM) part-of-speech (POS) tagger, that generates Penn Treebank POS tags for tokens. Articles were input as XML formatted documents (using the tags in the SGML format for LEADPARA, TEXT and HEADLINE), and output as XML files, with the identified parts of speech as the tags. To maximise tagger accuracy, the tagger was run using the *wsj-0-18-bidirectional-distsim.tagger* model, which enables fourth order bidirectional tag conditioning[3]. To match the input data formatting, the POS tagger *normalizeAmpersandEntity* option was set to 'true'.

The StanfordPOS tagger failed when encountering long tokens (such as '5p,4p10,4p9,4p7,4p6,4p3,4p,3p7,3p2,3p2'), forward slashes (as found in articles quoting currency exchange rates, such as: 'Austria/schilling, Belgium/franc'), lists of numbers (such as lists of fuel ratings for cars, or cinema listings), and numbers formatted with commas (such as population statistics: '5,826 7,688 1,329'). All the problem strings were one-off tokens, and together with the forward slashes, were not of interest for topic modelling, so each problem instance was replaced with a space character.

4.4 Document Restructuring based on POS Tags

New data sets were generated for select combinations of POS tags. The three sets of Penn Treebank tags of interest in this study were those tags for nouns, adjectives and verbs. The Penn Treebank uses multiple POS tags for each of these three parts-of-speech, to represent subclasses of these three items. These subclasses are not of interest in this study, and instead the tags of interest are grouped as indicated in Table 5.1. *Nouns* encompassed words assigned POS tags 'NN', 'NNS', 'NNP, and 'NNPS. Trials that included *adjectives* also contained POS tags 'JJ', 'JJR', and 'JJS'. Trials that included *verbs* also contained POS tags 'VB', 'VBD', 'VBG', 'VBN', 'VBP', and 'VBZ'.

The primary focus of this study was entities and concepts, so all trials contained nouns. Adjectives and verbs were included in some trials because adjectives can add detail to nouns, and verbs can inform on actions associated with nouns. While there are a large number of permutations around these POS that could be trialled, the select variants used in this study are summarised in Table 4.1, with examples included in Tables A.1, A.2 and A.3 in the Supplementary chapter.

Following the finding of Lau et al. (2014) that lemmatisation aided topic coherence, the restructured documents were lemmatised before up-weighting particular parts of speech. Lemmatisation was performed using the *morphy* software from NLTK[4], version 2.0.4, and was applied using the POS tag identified for each word. The *morphy* function reduced words to their base form, such as changing 'leveraged' to 'leverage', and 'mice'

[3]http:nlp.stanford.edusoftwarepos-tagger-faq.shtml#h

[4]http://www.nltk.org/howto/wordnet.html

to 'mouse'. Additionally, all punctuation was removed from the text, by removing any characters defined in Python's *string.punctuation*.

Three sets of trials (marked 'POS' in Table 4.1) reduced the original data set to only the select POS in the restructured documents, discarding all other word tokens. The remaining trials (marked 'All' in the *Text* column) up-weighted the select POS within the original text, by the factor shown in the *Nouns*, *Adj* and *Verbs* columns in Table 4.1. It was originally planned to also try five-fold weighting of the POS of interest, however, preliminary trials indicated such trials with higher numbers of topics generated memory stack errors when topic modelling, and since the triple weighting produced no major advantage, the trials for five-fold weighting of select POS were dropped. The weighting of words was done inline, rather than duplicating the weighted words at the end of the article. While some text analysis uses term frequency–inverse document frequency (tf-idf) to down weight common words, such as 'bill' in a corpus of government legislation, this technique was not used in this study as the SJMN corpus was broad in the subject matter, and its vocabulary, and the desire was to capture common content in the topics.

4.5 Named Entity Recognition

Named entity tags were generated using the Stanford Named Entity Recognizer (StanfordNER) (Finkel et al., 2005), version v3.3.1(2014-01-04), which is a Java implementation of a linear chain Conditional Random Field (CRF) sequence model. The StanfordNER was configured to produce 4 categories of named entities: *Location*, *Person*, *Organisation*, and *Miscellaneous*, using the *english.conll.4class.distsim.crf.ser.gz* classifier [5], trained from CoNLL 2003 Shared Task training data. The XML output was processed through a Python script to extract all items tagged as entities, and where the entity consisted of multiple words (e.g. 'San' 'Jose' or 'United' 'Airlines'), these tokens were concatenated to form a single token for topic modelling (e.g. 'SanJose' or 'UnitedAirlines'). The goal was to model 'SanJose' and 'SanFrancisco' as individual entities, rather than the bag-of-words approach of modelling 'Jose', 'Francisco' and 'San' separately. Three data sets were created based on entities, as shown in Table 4.1, including one containing only named entities.

4.6 Topic Modelling

Topic modelling was performed using the Mallet software from the University of Massachusetts Amherst (McCallum, 2002). The Mallet software was run to generate topics using the Latent Dirichlet Allocation (LDA) algorithm, converting all text to lowercase, to model individual features, not n-grams, and to remove words predefined in the Mallet English stop-word list prior to topic modelling. Each model was trained to use a hyper-parameter optimise-interval of 20, and 1000 Gibbs sampling iterations (i.e. the defaults settings). This study modelled a variety of settings for the *number of topics* parameter, in line with the recommendation of Claeskens & Hjort (2008) to determine parameter

[5] See the Stanford NER README.txt for details

settings by using various settings and choosing the model that performs best on a given metric (or metrics). Topic models were generated for 20, 50, 100, 200 topics (all as in Blei et al. (2003)) and also the more fine grained 500 topics, for each data set listed in Table 4.1. The goal was to have a set of coarse grained topics (i.e. smaller numbers of topics), and more fine grained topics (i.e. larger numbers of topics). Due to variability in the results on the word intrusion and observed coherence test evaluations, each trial was run ten times. Therefore, for each of the 22 data sets, at 20, 50, 100, 200 and 500 topic levels, there were 10 topics models generated for each level. While other studies have examined more topics on the SJMN database (for example Wei & Croft (2006) generated 800 topics), the weighting of parts-of-speech by duplicating words used in this current study made larger numbers of topics extremely time consuming to generate (as the more topics, the longer the time taken to generate the topics). As the results indicated a decline in performance at 500 topics compared to 200 topics on the metrics assessed, there appeared no benefit to generating more than 500 topics, and 500 topics was the maximum considered.

4.7 Evaluating the Semantic Coherence of Topics

The study by Lau et al. (2014) developed two measures of semantic coherence that were found to be well correlated with human evaluations of whether the top ten words of a topic combined to form an interpretable, coherent theme. While human evaluation of topic coherence is a gold standard measure for assessing coherence, such a manual evaluation was outside the scope of this current study. The two measures examined in the Lau et al. (2014) study were an observed coherence (OC) measure and an automated word intrusion (WI) task, and the authors have published open source software for these measures[6], software that was used in this study on document restructuring.

For the OC and the WI tests, the software provided a choice of methods for calculating word co-occurrence. The normalised point-wise mutual information (NPMI) method was chosen for this study, following the findings in Lau et al. (2014) that the NPMI method (developed in Bouma (2009)) is the least susceptible to bias for lower frequency words, and yielded more consistent results on topic level evaluations.

The two tests also required a reference corpus on which to base word co-occurrence assessments. Human evaluations of topic coherence draw on the vast bank of human experience to judge how likely words are to co-occur, but automated evaluations need to rely a specific reference corpus. Newman et al. (2010) suggested that an external reference corpus removes the risk of noise or idiosyncrasies related to the topic modelled data from clouding assessments of what co-occurrences are likely in common usage, and which are not. However, this raises the problem of what reference corpus available today can provide information on what word co-occurrences were likely in 1991, in San Jose. Studies such as Newman et al. (2010), Aletras & Stevenson (2013) and Lau et al. (2014) use Wikipedia, but there is a timeliness to Wikipedia that may not reflect the likelihood

[6]The version of these two evaluations used in this study was downloaded on the 1/5/2014 from https://github.com/jhlau/topic_interpretability

of particular co-occurrences from 1991. Following Mimno et al. (2011)'s findings that it is not inappropriate to use the training data set as a reference corpora, this study used the original SJMN data[7] as the reference corpus for the topic coherence tests. A second reference corpus of 42,652 Wall Street Journal (WSJ) articles from 1991, also obtained from the Tipster corpus, was used to confirm any significant differences identified in the pairwise comparisons of coherence scores. These 42,652 articles were grouped into 254 files, based on the groupings of the compressed corpus. The WSJ text was unaltered, not converted to lower case, and not lemmatised.

The OC and WI tests evaluated the top ten words for each of the 8,700 topics[8]. An NPMI OC score closer to 1 reflected greater co-occurrence, whereas a score of 0 indicated the words were independent. The WI evaluation required an intruder word to be inserted into a random location in each topic, and then the WI software used the word co-occurrence statistics from the reference corpus to choose which word was most likely to be the intruder. The intruder words needed to be words common to the corpus, but not related to the themes in the individual topic. The WI software rated accuracy in detection, either as '[1.0]' indicating the intruder word was automatically detected, or '[0.0]' that the intruder word was not detected by the automatic evaluation. A supplemental Python script was written for this study to determine the proportion of all topics where the WI software automatically detected the intruder word, expressing this as a value between 0 and 1, where 1 meant 100% of all intruder words were detected. This proportion was recorded for each topic model generated in this study.

The R package was used to calculate all descriptive statistics. The trials were compared using Tukey all-pairwise multiple comparisons over linear mixed effect models, generated using the R package nlme[9], and compared via the multcomp[10] and factorplot[11] R packages, and a family-wise confidence level of 95% was used for each comparison. The Tukey all-pairwise comparison method was chosen due to the imbalance in sample sizes between the different *number of topics* settings (Foster et al., 2006).

4.8 Document to Topic Level Evaluations

4.8.1 Information Retrieval

An information retrieval (IR) task was used to assess whether the restructuring of the source documents had improved or degraded performance on such a retrieval task compared to the baseline. Baseline performance was taken as the performance on the IR task of topics generated from the original, unaltered data set. The IR task employed

[7] Note, the reference corpus included only the headline, lead and main tagged text from the SJMN SGML files of the 90,257 articles

[8] 8700 topics: 10 runs generating 20 topics each run, 10 runs generating 50 topics each, 10 runs generating 100 topics each, 10 runs generating 200 topics each, and 10 runs generating 500 topics each

[9] http://cran.r-project.org/web/packages/nlme/index.html

[10] http://cran.r-project.org/web/packages/multcomp/index.html

[11] http://cran.r-project.org/web/packages/factorplot/index.html

followed Wei & Croft (2006) and Wang et al. (2007), to use topics for query expansion in ad hoc information retrieval. Under this assessment, an improvement in performance is represented by a higher calculated probability. The evaluation was trialled using the SJMN 1991 TREC data set, queries 51-150, both an unlemmatised and also a lemmatised version of this query set, where lemmatisation altered 31 of the 94 query entries (for example, changing 'leveraged' to 'leverage'). The software to implement this IR task was provided by Z. Zhao (personal communication, 15th August, 2014), and this software considered the top 1000 ranked documents in its calculations.

4.8.2 Document - Topic Association

The document to topic link was also reviewed by considering the topic with which the document has the strongest primary association. The goal was to determine if restructuring produced stronger associations between documents and their primary topics. Mallet can produce a composition file that indicates the degree of association between each generated topic and each individual document as a score between 0 and 1, via the *–output-doc-topics* option. For example, article sjmn91-06210046 is most strongly associated with topic 358, and the topic model estimates the contribution of topic 358 to this document at 0.49 (or 49%). From such output, using a Python script, each document's strongest degree of association was extracted, the scores were binned using ten bins of width 0.1, and a document count was accumulated for each bin (e.g. bin {0.4 -.49} would be incremented by 1 for article sjmn91-06210046). Linear mixed effects models were generated by bin, with a fixed effect for the trial and a random effect for the individual run, and contrasted using Tukey all-pairwise multiple comparisons via the R multcomp package, for a 95% significance level.

4.8.3 Topic-Descriptor Alignment

The SJMN corpus is provided with a set of manually annotated descriptor tags, where tags represent themes that the annotators judged to be appropriate for individual articles. Of interest was whether restructuring the documents improved or degraded any alignment of topics to sets of descriptor tags. To assess this, a Python script was created that built a list of all descriptors from all documents that had each topic as one of its top three topics. Average counts of descriptors per topic were compared across trials using a Wilcoxon rank sum test in the R statistical package. The results for this, and all other evaluations are included in the Results chapter.

TABLE 4.1: *Document Structures by Trial*

Trial Identifier	Lemma	Text	Weight			
			Nouns	Adj	Verbs	Lead Paragraph
Original*	N	All				
Lemmatised	Y	All				
Nouns-RL-U	Y	POS	x1			
Nouns+Adj-RL-U	Y	POS	x1	x1		
Nouns+Adj+Verbs-RL-U	Y	POS	x1	x1	x1	
Double Nouns-WL-U	Y	All	x2			
Double Nouns+Adj-WL-U	Y	All	x2	x2		
Double Nouns+Adj+Verbs-WL-U	Y	All	x2	x2	x2	
Triple Nouns-WL-U	Y	All	x3			
Triple Nouns+Adj-WL-U	Y	All	x3	x3		
Triple Nouns+Adj+Verbs-WL-U	Y	All	x3	x3	x3	
Double Nouns-WL-F	Y	All	x2			Five-fold
Double Nouns+Adj-WL-F	Y	All	x2	x2		Five-fold
Double Nouns+Adj+Verbs-WL-F	Y	All	x2	x2	x2	Five-fold
Triple Nouns-WL-F	Y	All	x3			Five-fold
Triple Nouns+Adj-WL-F	Y	All	x3	x3		Five-fold
Triple Nouns+Adj+Verbs-WL-F	Y	All	x3	x3	x3	Five-fold

Trial Identifier	Lemma	Text	NER
NE-R-U	N	NE only	
NE-W-U	N	All	Unify NE elements
NE-Double-W-U	N	All	Add unified NE reference to original text

* Baseline: Original text without any weighting or lemmatisation
In the above table:

'RL'= POS lemmatised and reduce articles to only these select POS
'WL'=POS lemmatise and weight these select POS within the original text
'x2' = Double weight; 'x3' = Triple weight
'Five-fold'= x 5 weight in text
'-U'=Uniform weight to all paragraphs
'-F'=Five-fold weight selected POS in lead paragraph
'NER'=Named Entity
'Unify NE' = Retokenise multi-word entities. E.g. 'San'+'Jose'='sanjose'

Chapter 5

Results

5.1 Introduction

To determine if restructuring documents enabled more semantically coherent, valid and reliable topics to be generated, this study used a mix of topic level evaluations and document to topic evaluations. The document level evaluations were the Observed Coherence (OC) and Word Intrusion (WI) tests developed in Lau et al. (2014). The aim was to determine if any of the restructuring trialled in this study produced higher OC scores or WI detections, indicating more semantically coherent topics, and/or reduced the variance of such scores as a sign that restructuring more reliably generated coherent topics. The document to topic level evaluations sought to determine if restructuring improved or degraded performance on an information retrieval (IR) task, or improved or degraded the average associations between the source documents and each document's primary (most relevant) topic. Maintaining the same performance as the baseline was taken as an indication that the topics from the restructured data were no worse than the topics generated against the baseline in representing the content of the documents, as a rough measure of validity. While these measures are imperfect, and human evaluation would be the best measure of the validity, reliability and interpretability, the number of trials and number of topics generated made human evaluation infeasible. The results from these approximate measures are detailed in this chapter, and discussed in the next chapter. Before presenting these results, this chapter will first review the San Jose Mercury data set, to provide some context to the complexity and variability found in these evaluations.

TABLE 5.1: *Summary of SJMN Word Tokens for Select Parts-of-Speech*

POS	Number of Instances	Percentage of Total Instances	Number of Distinct Word*-POS Pairs	Penn Treebank Tags
Common Nouns	7,640,668	21%	84,669	NN, NNS
Verbs	5,560,451	15%	40,833	VB,VBD,VBG,VBN,VBP,VBZ
Proper Nouns	5,261,088	14%	160,773	NNP, NNPS
Adjectives	2,586,019	7%	77,139	JJ,JJR, JJS
Other	14,779,316	43%	209,772	
TOTAL	35,827,542		573,186	

Note: A *Word Token to Part-of-Speech* pair would be : 'learn' tagged as a common noun (NN)

TABLE 5.2: *Summary of SJMN Articles by Newspaper Section*

Section	No. of Articles	Section	No. of Articles	Section	No. of Articles	Section	No. of Articles
front	24181	extra	1003	computing	526	westextra	251
local	12400	arts&books	976	weekend	521	specialsection	230
sports	11873	home	948	peninsulahome	513	eastextrapart	190
business	10624	businessmonday	905	science&medicine	502	tvmagazine	163
living	7514	travel	818	theweeklypart	448	professionalcareers	157
californianews	3231	peninsulaextra	777	theweekly	444	stanfordextra	67
editorial	2941	perspective	646	venture	443	bq	2
generalnews	2129	eastextra	616	religion&ethics	423	ge	1
eye	1082	garden	543	drive	334		
food	1038	west	530	theletterspage	267	Total:	90,257

Note: Calculated using the SGML <SECTION> tags provided with each SJMN article

TABLE 5.3: *Examples of SJMN Manually Annotated Descriptor Tags*

professional (8830)	us (8635)	chart (5199)	brief (4927)	company (4693)
result (4095)	san-jose (3688)	baseball (3595)	list (3157)	war (3001)
mideast (2760)	college (2715)	california (2617)	change (2572)	government (2477)
death (2392)	controversy (2297)	sport (2197)	profile (2185)	official (2147)
opinion (2105)	end (2079)	president (2053)	san-francisco (2034)	letter (2018)
softball (40)	telecommunication (39)	north (39)	fresno (39)	hollister (39)
drawing (1)	fishing (1)	gridley (1)	ski (1)	fate (1)

Note: Calculated using the SGML <DESCRIPT> tags provided with each SJMN article. The number of articles linked to each tag is shown in parentheses.

5.2 Data Set

5.2.1 Original Data Set

The corpus of San Jose Mercury newspaper (SJMN) articles had an extensive vocabulary, discussing a broad set of entities and themes from a variety of perspectives. The original SJMN corpus consisted of 90,257 articles, together containing over 35 million word tokens, as indicated in Table 5.1, which summarises tokens by part of speech (POS). Approximately 20% of the articles (18,188/90,257) contained 100 words or less, and only 4% of articles (4,080/90,257) exceeded 1,000 words, with an average of 396 tokens per article (for the article header, lead and main text combined).

The articles in the SJMN corpus are drawn from 39 newspaper sections (detailed in Table 5.2), suggesting a range of perspectives or focus of the articles. The manually assigned annotations identified 1,617 distinct themes in the articles (a sample of these descriptive annotation tags are shown in Table 5.3). The breadth of the sections and annotated themes in the corpus suggest that the SJMN corpus covers a broad array of subject material, from a variety of different perspectives.

5.2.2 Data Preprocessing

The word counts of the restructured data sets are shown in Table 5.4. Prior to generating these data sets, problematic long strings that caused memory violations during tagging

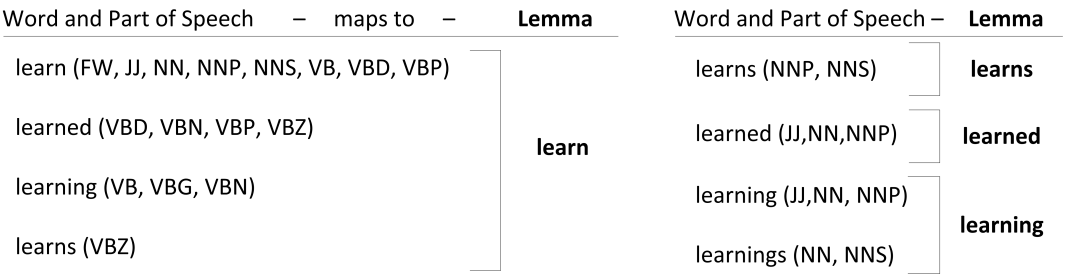


FIGURE 5.1: Example of different word to lemma mappings for variants of 'learn'

were removed from 58 articles, in line with the discussion in the previous chapter.

The intention behind lemmatising the text had been to reduce the data set to a smaller set of lemmas by removing tense and declension. However, this reduction was not substantial with 298,749 distinct word tokens being reduced to 289,066 distinct lemmas. The lack of substantial reduction was due to lemmatisation being set to use POS, meaning that a word such as 'hearings' lemmatised to 'hear' for verbs, to 'hearing' for common nouns, and to 'hearings' for proper nouns, producing three lemmas from one initial word token. There were approximately 15,200 word tokens that mapped to different lemmas depending on the POS for any given instance of that token.

While the same word (e.g. 'hearing') could lemmatise to different lemmas (e.g. 'hear', 'hearing', 'hearings'), it was also the case that individual lemmas were produced by different words. Figure 5.1 indicates how the lemma 'learn' is generated by multiple words, for various parts of speech, and how the word 'learning' can be lemmatised to 'learn' or 'learning' depending on part of speech.

The process of lemmatising increased the size of the data set from 35.4 million tokens to 36.2 million tokens, and slightly increased the average time to topic model the lemmatised data set over the baseline set, as indicated in Figure 5.5. This increase was the result of word tokens being parsed during POS tagging, prior to lemmatising. For example, a single token '(text)' was split to three: '(','text', ')'. The parsing would be done by the Mallet topic modelling tool anyway[1], and the inflation of tokens is a timing issue due to the sequential, unintegrated, process of POS tagging and then topic modelling used in this study.

5.3 Coherence

Each topic model was generated ten times for each of the five *number of topic* settings, for each data set detailed in Table 5.4. Heteroscedasticity was observed between the different settings for *number of topics*, observed both in Figures 5.2 and 5.3, and found when

[1]Note, all trials were configured to disregard punctuation when topic modelling

TABLE 5.4: *Word Statistics by Trial*
(rounded to nearest thousand)

	Word Counts	
	Paragraph Weights	
Weight	Uniform	Double Lead
Original*	35,462,000	
Lemmatised	36,190,000	
Nouns-RL-U	12,902,000	21,791,000
Nouns+Adj-RL-U	15,488,000	26,114,000
Nouns+Adj+Verbs-RL-U	22,398,000	37,084,000
Double Nouns-WL-U	49,092,000	57,981,000
Double Nouns+Adj-WL-U	51,678,000	62,304,000
Double Nouns+Adj+Verbs-WL-U	58,588,000	73,275,000
Triple Nouns-WL-U	62,355,000	80,393,000
Triple Nouns+Adj-WL-U	67,166,000	88,418,000
Triple Nouns+Adj+Verbs-WL-U	80,986,000	110,359,000
NE-R-U	2,570,775	
NE-W-U	40,404,032	
Double NE-W-U	42,974,807	

Note: 'RL' : topic model select POS only; 'WL' : topic model with select POS up-weighted in original text

generating non-linear mixed models[2] with a fixed effect for *trial* and *topic*, and a random effect for the individual run. As a result, all subsequent evaluations and comparisons were done separately for each *number of topics* level (e.g. trials compared at the 500 topics level, and separately compare trials at 200 topics level, and so on). At this level, QQplots indicated each set of scores were approximately normally distributed, for each test, for each data set, for each separate *number of topics* setting.

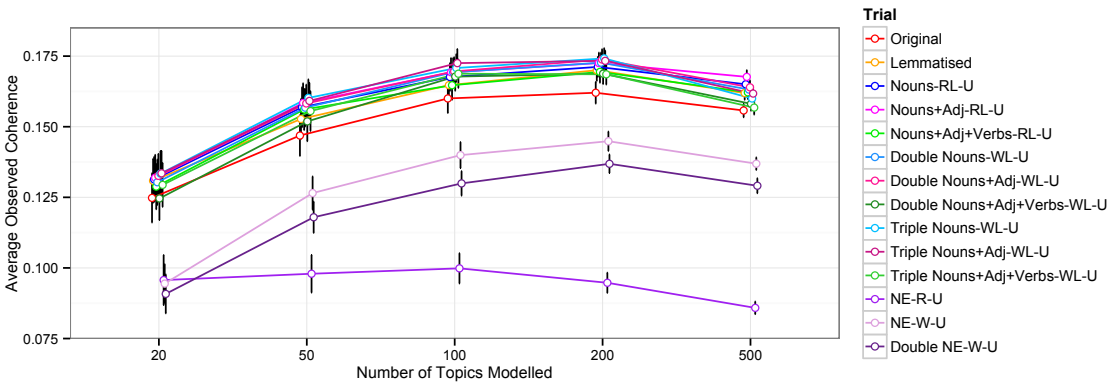
Non-linear models were generated for the OC scores and WI detection proportions respectively, with *trial* as the fixed effect and the *run* as a random effect. Tukey pairwise contrasts to compare the different trials produced a large number of comparisons[3], which are summarised as factor plot graphs[4] in Figure 5.4. It should be noted that while the NE data set was re-tokenised for multi-termed named entities (affecting approximately 3% of the corpus)[5] the reference corpus was not. Therefore, no comparison should be attempted between the POS trials and the NE trials. The NE trials are included for completeness, but as will be detailed in section 5.7 and discussed in section 6.5 in the next chapter, the simplistic nature of NE handling used in this study did not resolve different references to the same NE (e.g. did not resolve 'Jane Smith' and 'Jane' to the same entity). While the method was sufficient for consistently named geographical locations such as *Denver* and even multi-termed locations names such as *Los Angeles*, overall the review of the NE topics produced indicated it was inadequate for assessing the true frequency of references

[2]Generated using the R package *nlme*

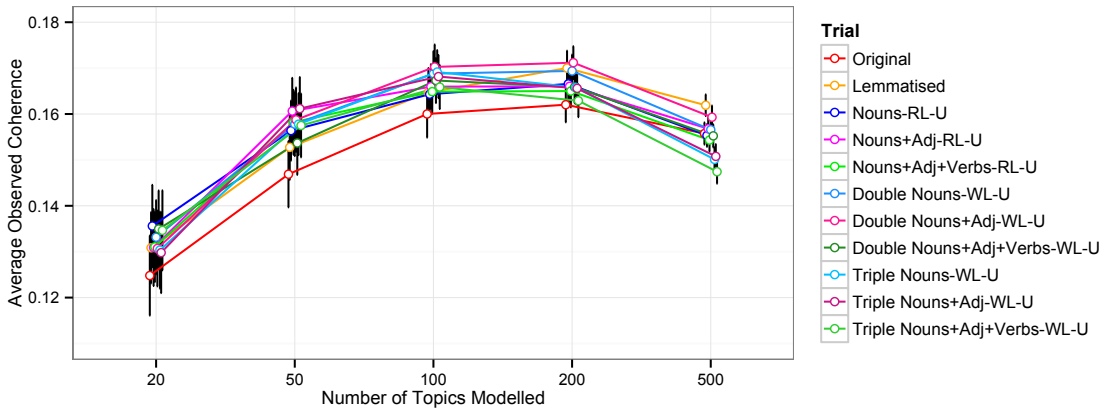
[3]253 comparisons for each of the five *number of topics* settings

[4]Factor plots generated using the *factorplot* R package

[5]Out of the over 35 million word tokens, 4,190,124 entities (people, places, locations, etc.) were identified, and 1,146,516 of these entities where re-tokenised to form multi-termed entities (e.g. *PeterPan*)



(a) Uniform weight paragraphs

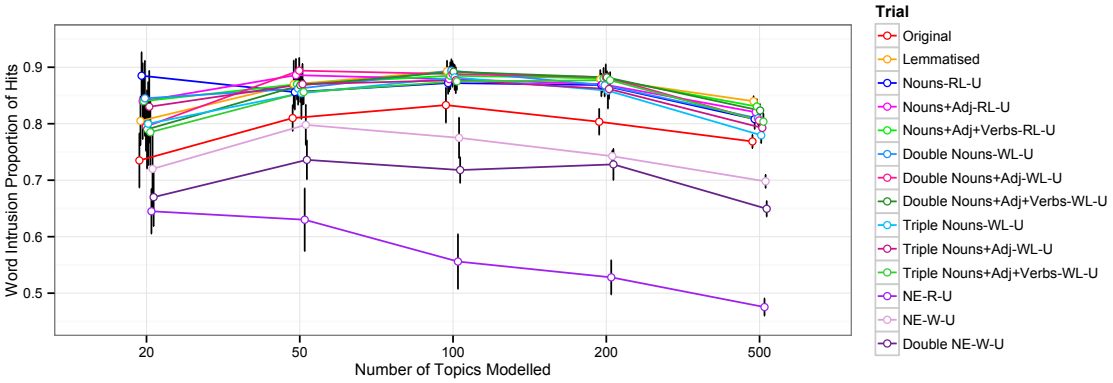


(b) Double weight lead paragraph

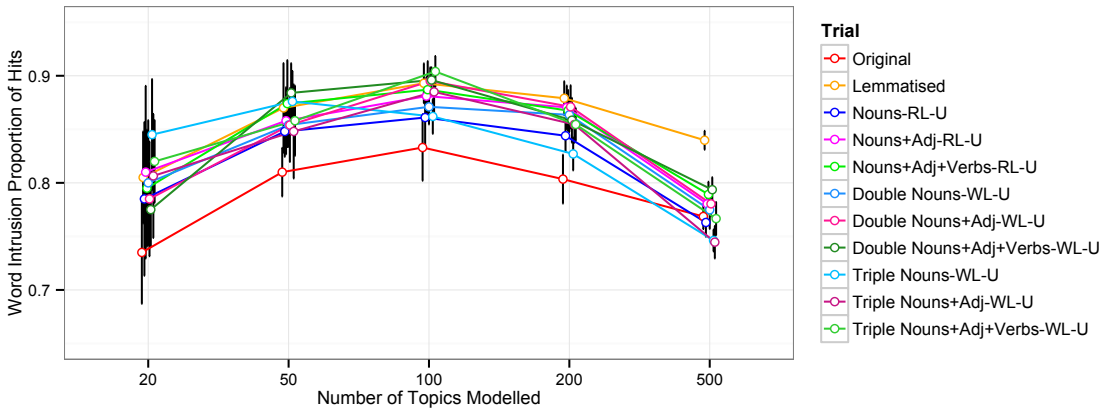
FIGURE 5.2: Observed Coherence (OC) scores, by *number of topics* and trial. The mean and standard deviation values are detailed in Table A.4 in the Supplemental chapter.

to a given entity, as discussed in the qualitative analysis of topics below.

Figures 5.2 and 5.3 suggest the POS trials produced OC score means above the baseline, however, only when the number of topics generated was 200 or 500 were any such increases for the POS trials statistically significant compared to the baseline (p<.05). Reducing the data set to nouns and adjectives produced significantly better than the baseline OC scores for both 200 and 500 topics, as did doubling nouns with adjectives in the original text. In addition, doubling only the nouns within the original data set also significantly improved the OC scores (p<.05) at both the 200 and 500 topics level. The factor plots in Figure 5.4 indicate that when generating 500 topics, the OC scores benefited from the data set being reduced to the select POS of interest, this was true for nouns alone, or with adjectives or with adjectives and verbs. Triple weighting select POS at 500 topics did not improve performance against the baseline, and reduced performance compared to the other forms of restructuring trialled (p<.05). In contrast, at 200 topics,



(a) Uniform weighting of paragraphs

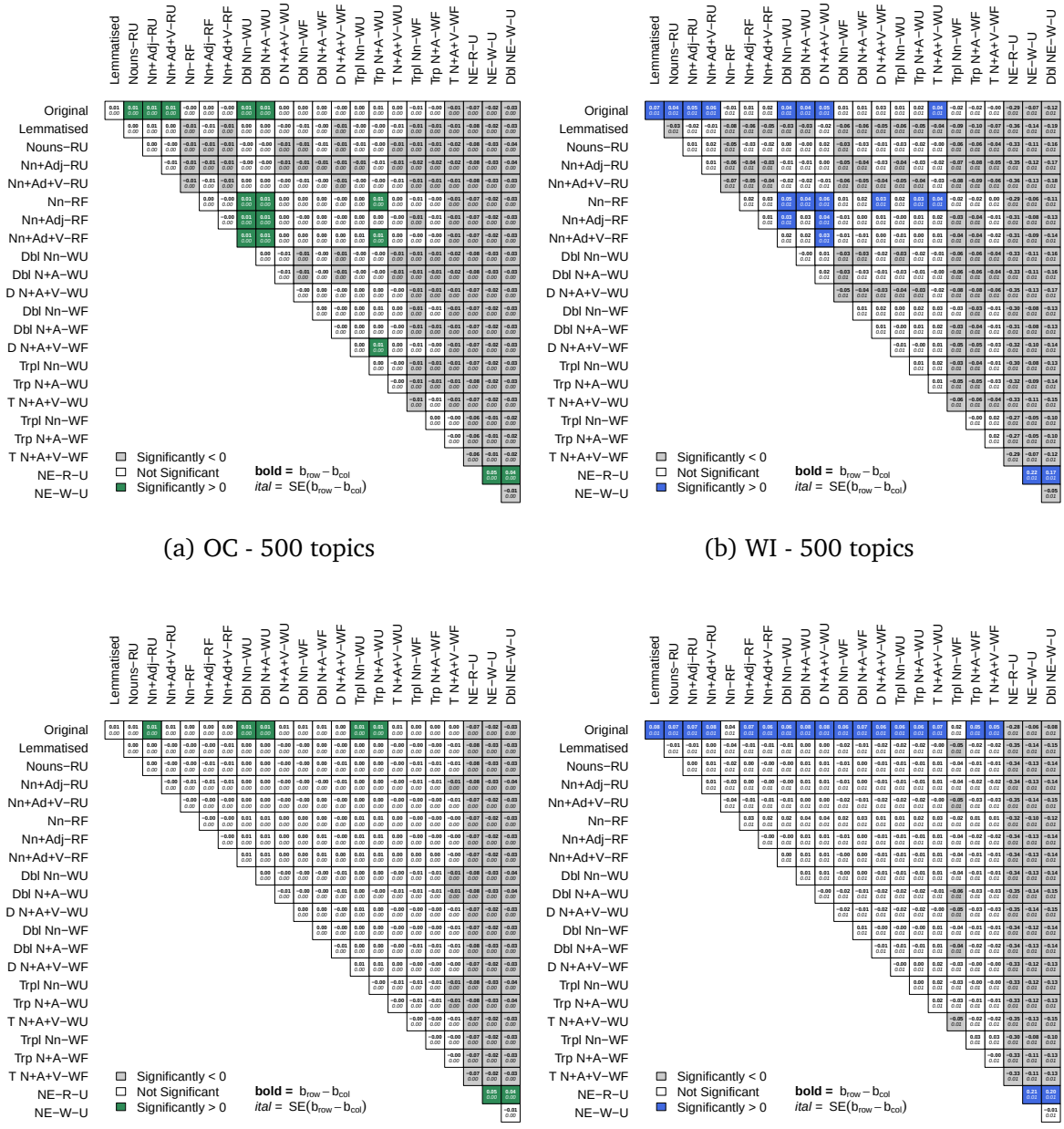


(b) Double weight lead paragraph

FIGURE 5.3: Proportion of trials where intruder word was successfully detected by *number of topics* and trial. The mean and standard deviation values are detailed in Table A.5 in the Supplemental chapter.

the OC scores benefited from triple and double weighting nouns, or nouns with adjectives ($p<.05$ for both), but the inclusion of verbs removed any such gains. The mean and standard deviation values plotted in Figures 5.2 and 5.3 are provided in Tables A.4 and A.5 in the Supplemental chapter.

For the Word Intrusion test, Figure 5.4 shows that restructuring produced improved detection rates across all *number of topic* settings, improvements that were statistically significant ($p<.05$), however, there was no form of document restructuring that was consistently better for all *number of topic* settings. Doubling nouns and adjectives within the original text improved performance on and above 50 topics, simply lemmatising text improved word intrusion detection at 100 topics and above, and all POS weighted trials with uniform paragraph weighting performed better than the baseline for 200 topics. However, performance for some of these variants was no better than the baseline at 500 topics. As was found with the OC tests, replacing the original text with select POS produced improved performance against the baseline at both the 200 and 500 topic models.



(a) OC - 500 topics

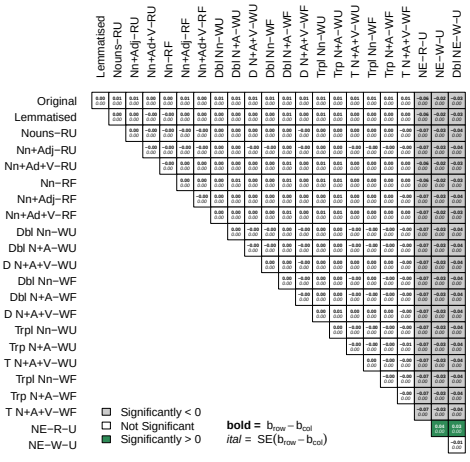
(b) WI - 500 topics

(c) OC - 200 topics

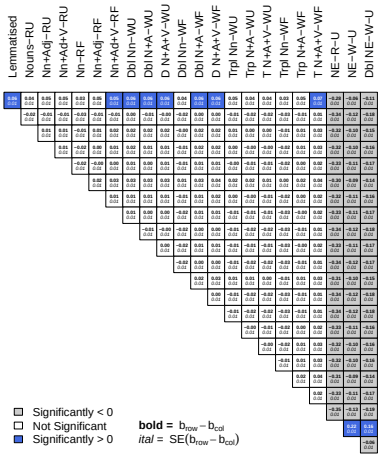
(d) WI - 200 topics

FIGURE 5.4: Factor plot graphs depicting Tukey pairwise comparisons by trial and *number of topics* for the Observed Coherence (OC) scores and Word Intrusion (WI) successful detection proportions (500 and 200 topics)

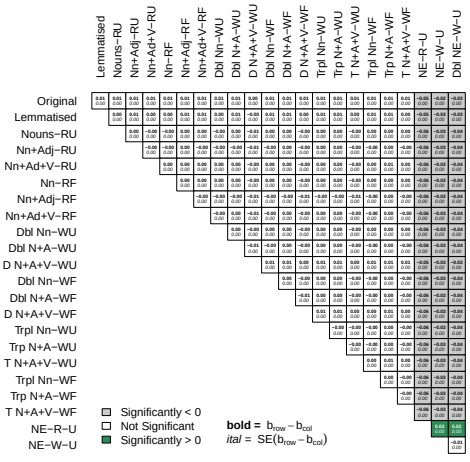
Note: The green and blue squares indicate the mean of the trial listed on the x-axis is significantly greater than the mean for the trial on the y-axis, for the respective cells (for a 95% confidence level). The grey cells indicate the reverse, that the mean of the trial on the x-axis are significantly less that the mean of the trial on the y-axis for the respective cell (with 95% confidence).



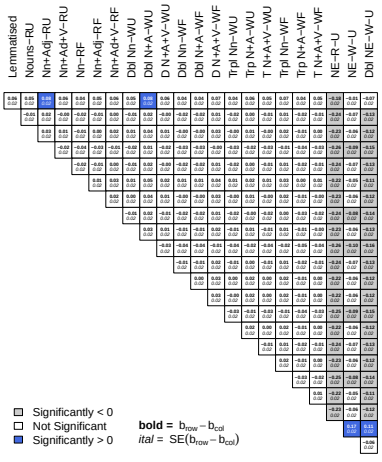
(e) OC - 100 topics



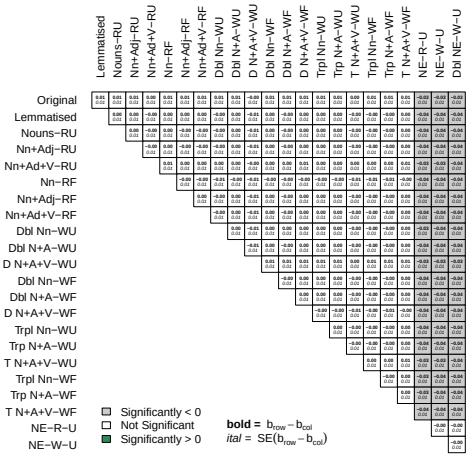
(f) WI - 100 topics



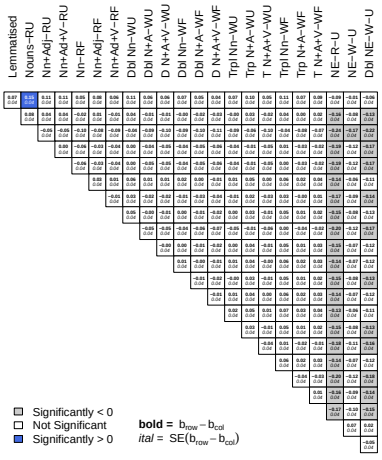
(g) OC - 50 topics



(h) WI - 50 topics



(i) OC - 20 topics



(j) WI - 20 topics

FIGURE 5.4: Factor plots showing Tukey pairwise comparisons between trials, by number of topics, for Observed Coherence (OC) scores and Word Intrusion (WI) successful detection proportions (100, 50 and 20 topics)

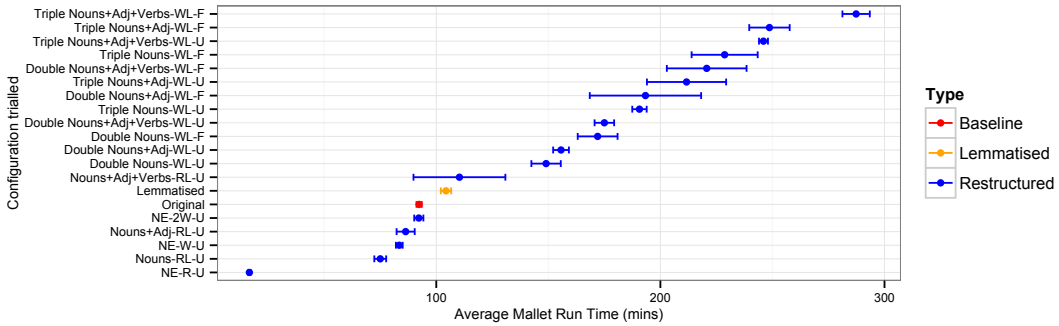


FIGURE 5.5: Average time to run topic modelling by trial for a 200 topic model (includes both the time to import data into the Mallet format and the time to train the topic model)

Weighting the lead paragraph produced no observable advantage to uniformly weighting all paragraphs, and at the 500 topic level, weighting the lead paragraph produced less coherent topics than uniform weighting of paragraphs.

The improvements in topic coherence for these select models could not be confirmed using a reference corpus of 1991 Wall Street Journal (WSJ) articles. Density plots and QQPlots indicated the OC scores were positively skewed, with over 80% of the coherence scores falling below 0.1. In addition, the WI detection rates were much lower than where the SJMN reference corpus was used (WSJ: M=0.29, SD=0.07 ; SJMN: M=0.82, SD=0.09). Density plots, QQplots and box plots by number of topics that indicate these patterns are included in Figure A.1 in the Supplemental chapter.

5.4 Topic Modelling Run Times

As shown in Figure 5.5, reducing the data set to only named entities or to select parts-of-speech reduced the average time to generate topic models. Likewise, consolidating multi-termed NE to single terms reduced topic modelling times. Both the NE tagging and the POS tagging took less than one second per article, respectively[6].

5.5 Information Retrieval

The information retrieval (IR) evaluation indicated that neither lemmatising alone or with POS based restructuring changed the mean average precision score (MAP) from the baseline of 0.1903, even if the queries were also lemmatised. The trial in which the original text was replaced with only named entities produced an improved MAP of 0.2198, at $\lambda = 0.5$, which was unexpected given the limited word set in the restructured documents. Retokenising NE in the original text increased the MAP to 0.1990, at $\lambda = 0.6$, against the lemmatised queries even though the NE data set was not lemmatised.

[6]Detailed timings for topic modelling (by trial) can be found in Table A.6, POS tagging in Table A.7 and NE in Table A.8 in the Supplemental chapter

TABLE 5.5: *Average Number of Documents by Document-Primary Topic Association Strength*

	Document to Topic Association [0-1]							
	0.9+		0.8-.89		0.7-.79		0.7 and above	
Trial	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Original	1,355	(91)	2,108	(137)	2,833	(122)	6,296	(209)
Lemmatised	1,223	(77)	2,159	(232)	2,706	(210)	6,087	(402)
Nouns-RU	1,667	(69) *	3,199	(202) *	4,157	(286) *	9,023	(483) *
Noun+Adj-R-U	1,487	(117)	2,811	(143) *	3,586	(176) *	7,883	(379) *
Noun+Adj+V-R-U	1,261	(141)	2,231	(219)	2,735	(180)	6,227	(471)
Dbl Noun-W-U	1,184	(112)	2,288	(170)	2,710	(220)	6,182	(413)
Dbl Noun+Adj-W-U	1,361	(148)	2,287	(165)	2,721	(196)	6,368	(393)
Dbl Noun+Adj+V-W-U	1,374	(154)	2,254	(248)	2,583	(310)	6,211	(631)
Trpl Noun-W-U	1,084	(205) *	2,225	(233)	2,521	(166)	6,111	(300)
Trpl Noun+Adj-W-U	1,209	(139)	2,262	(257)	2,640	(133)	5,830	(420)
NE-R-U	12,069	(194) *	12,431	(74) *	10,471	(97) *	34,971	(229) *
NE-W-U	1,145	(49) *	2,285	(228)	3,063	(182)	6,493	(376)
Dbl NE-W-U	1,739	(106) *	2,543	(226) *	3,615	(182) *	7,897	(364) *

* Significantly different from the baseline.
Note: Statistical significance evaluated using Tukey all-pairwise multiple comparisons with a 95% confidence interval, calculated for each individual bin count

Counts for bins 0.7 to 0 are provided in Table A.9 in the Supplemental chapter

Re-tokenising NE's and double weighting those NE in the original text produced a MAP of 0.2108 at $\lambda = 0.6$. Therefore, performance was not degraded by restructuring the source documents.

5.6 Other Tests

Restructuring the documents did not improve the alignment between the manually annotated descriptors and the generated topics. In reviewing the 200 topic model, across the 10 versions run for each trial (n=2000 per trial), the mean number of descriptors linked to each topic in the original model was 386 (SD=138, ranging from 26 to 773 descriptors per topic). When compared using a Wilcoxon rank sum test, none of the restructured document data sets were statistically different to this baseline ($p<0.05$).

Compared to the baseline, four forms of document restructuring produced a statistically significant increase in the number of documents strongly associated to their primary topic (for associations above 0.7, $p<0.05$). Table 5.5 shows that two of these trials were those that focused on named entities, with a substantial increase for the trial that replaced the original text with only named entities in each article. This was despite this trial performing poorly on the coherence evaluations. Improvement was also found in the trials that reduced the data set to only nouns, or only nouns and adjectives.

5.7 Qualitative Analysis

Examining topic differences between each of the ten runs, it was clear that the words in the topics varied substantially from run to run, for a given trial. For example, in the 100 topic models, for a data set restructured to only include nouns, all of the ten runs had one

topic containing the word *wine*[7]. However, in one run the topic included {*water wine plant tree garden flower fruit valley drought soil winery california leaf seed gallon*}, while another run produced {*wine croatia yugoslavia army winery republic yugoslav slovenia serbia valley fruit cabernet chardonnay sauvignon serbs*}. In the other eight runs over this data set, the wine topics included variants of *winery valley fruit beer bottle cabernet chardonnay sauvignon napa alcohol vineyard grape*. While this latter combination occurred in eight out of ten runs, there is no indication from any single run that a link of *wine* to *Croatia* is less likely than linking *wine* to *bottle* were the topic model to be regenerated. This example indicates that if a researcher seeks to interpret topics directly, then caution is needed when seeking to infer meaning in the association of the top ten to twenty words in a topic.

As would be expected, restructuring the source documents to reduce each article to only NE produced a different set of topics. Rather than topics for general themes like *wine*, topics instead consisted of entities, people, places and organisations. However, a review of these topics indicated that reducing the corpus to only NE produced what appeared to be spurious links. For example, while *geologicalsurvey* was often placed in a topic with *richter* in multiple topic modelling runs, in one run it was placed in a topic with *buffett*. There were no articles in the SJMN corpus that mentioned both *geological survey* and *Buffett*, or even had a cross over relevant to the financier. It should be noted that this was one topic, in one of 10 runs of a 200 topic model, so a 1 in 2,000 topic occurrence. However, other runs[8] also produced strange links between *geologicalsurvey* and *StarTrek* in one run, *geologicalsurvey* with *Taylor* and *Burton* in another. Again, there were no commonalities in theme, people or place between these entities in any of the documents referencing *geological survey*. This appeared to be an instance of multiple unconnected themes being combined into a single topic, as discussed in Mimno et al. (2011).

5.8 Conclusion

This study found that the POS based restructuring of the source documents had more effect when generating 200 or 500 topics, while the highest OC scores and WI detection rates occurred for 100 and 200 topics. When generating 200 topics, the average OC scores increased when the source documents were reduced to only nouns and adjectives (M=0.173, SD=0.084) or only nouns (M=0.171, SD=0.081), compared to the baseline (M = 0.162, SD=0.087), both representing a statistically significant improvement over the baseline (p<.05). Average automated WI detection increased when reducing articles to nouns and adjectives (M=0.870, SD=0.019), or to only nouns (M=0.869, SD=0.028), which were both statistically significantly better (p <.05) than the baseline (M=0.803, SD=0.032). The 200 topic modelling run times improved if the data set was reduced to nouns and adjectives (M=86 mins, SD=4 mins), or to nouns only (M=75 mins, SD=3 mins), compared to the baseline (M=92 mins, SD=1 min) (significant at p <.05). While statistically better than the baseline, the difference between the nouns only, and the

[7]See Table A.10, lines 1-10, in the Supplemental chapter for the actual topics

[8]See Table A.10, lines 11-23 in the Supplemental chapter for the actual topics

nouns with adjective trials were not statistically significant ($p > .05$) on any of these three measures. The number of documents identified by Mallet as being over 70% related to their assigned primary topic was higher for nouns and adjectives (M=7,883, SD=379), and nouns only (M=9,023, SD=483), compared to the baseline (M=6,296 SD=209), with both models producing statistically significant ($p < .05$) improvements over the baseline.

These results suggest that lemmatising and reducing documents to only nouns, or nouns and adjectives, when 200 topics are generated, produces topics that contain words that are more likely to co-occur in the source articles. Such topics may represent more interpretable themes (e.g. a topic about 'wine production'), which would be more meaningful and useful if the goal was to directly review the top ten to twenty words of the topic to gain an insight into the themes contained in the source articles.

Chapter 6

Discussion

6.1 Introduction

From the models trialled, and the evaluations performed, it appears that there are advantages in the forms of restructuring documents that reduce noise in the dataset, forms such as lemmatisation and parts-of-speech (POS) based data reduction. This chapter will review the results for each of the main restructuring approaches considered, starting with a review of the effect of lemmatisation. Next, the weighting of the lead paragraphs will be discussed, followed by discussions on the weighting of POS, and on weighting the named entities. Finally, issues related to the particular evaluations, rather than particular document restructuring, will be discussed.

6.2 Lemmatisation

Lemmatising with reference to POS meant that the lemmatised tokens perhaps more accurately capture the sense in which a word was used, however such POS-lemmatising made the data set noisier than lemmatising without reference to POS. For example, consideration of POS meant that verb instances of 'learns' were converted and counted with other instances of 'learn', yet noun instances were not. This increased the frequency of 'learn' and reduced the frequency of 'learns', without eliminating 'learns' from the data set. It is not clear that separating 'learns' from 'learn', such that both could independently appear in topics, would necessarily be an advantage in a complex set of documents such as a newspaper corpus. It should be noted that the Lau et al. (2014) study used non-POS based lemmatisation, where all instances of 'learns' would be reduced to 'learn'.

Regardless of the difference in considering POS when lemmatising, this current study agrees with the conclusion of Lau et al. (2014) that lemmatising appeared to be important to the word intrusion evaluation. Lemmatisation produced an improvement in intruder word detection that was statistically significant ($p<0.05$) when the number of topics was 100, 200 or 500, though not for lower settings (i.e. 20, 50). Following the Lau et al. (2014) finding that such automated detection correlates well with human evaluations, and is suggestive of more coherent topics, the finding of significance in this study suggest lemmatising aids topic coherence at least for 100, 200 and 500 topics.

Lemmatisation produced no statistically significant differences to the baseline on the other evaluation measures considered in this study. The slight increase in topic modelling run time noted in the lemmatised set may not be an issue, however, if lemmatisation were integrated into the topic modelling process, rather than as a separate preprocessing step.

The conclusion suggested by these finding is that topic quality is not reduced, and may be improved, by POS based lemmatisation, at least in the aspects of topic coherence measured in the automated word intrusion detection task. However, lemmatisation alone does not appear to be enough to produce clear improvements in topic quality when assessed by a broader range of evaluation measures. Non-POS based lemmatisation may yield further improvements if it reduces noise in the source data through a coarser combination of word tokens (e.g. 'learns' always combined with 'learn').

6.3 Weight by Section

Weighting the lead paragraph over the remainder of the article provided no measurable benefit to improving topic quality, on any of the evaluations considered in this study. Such weighting increased the size of the data set, substantially slowed topic modelling run-times, and produced observed coherence scores and word detection rates often worse than the baseline, and worse that other trials without such weighting

It may be that in the newspaper corpus used in this study, the lead paragraph does not represent a summary of the word tokens that appear in subsequent paragraphs. This may be due to the length of the articles, the nature of making the lead paragraph in news attention grabbing, or that journalists are trained to avoid word repetition.

The conclusion suggested by this study is that up-weighting the lead paragraph in news articles is not useful to improve topic quality. It may be the case that weighting by section may be more productive in document types when there is a section that truly summarises the concepts (i.e. specific word tokens) expanded on in later sections, such as an abstract in research papers or patent applications.

6.4 Weight by Part-of-Speech

The original proposition considered was that nouns, in providing semantic content, may point to content based themes latent in documents and, in turn, can produce semantically coherent topics under LDA topic modelling. Nouns were modelled separately to other POS, and also were augmented by adjectives, or by adjectives and verbs in some trials.

The results of the evaluations conducted in this study indicate, firstly, that such approaches do not aid, but also do not reduce, topic coherence when the number of topics is low (less than 200). Secondly, up-weighting by doubling or tripling select POS within the text substantially slows the topic modelling run time, which can be problematic with large datasets. Thirdly, triple weighting did not aid performance, and in some instances reduced performance on the observed coherence (OC) and word intrusion (WI) tests compared to the gains in lemmatising the data. Fourthly, double weighting produced

some gains over the baseline in the topic level OC and WI tests, but not in the other evaluation measures. Fifthly, reducing the dataset to nouns, or nouns with adjectives, produced improvements in topic OC that were statistically significant over the baseline ($p<.05$) for the 200 and 500 topic models. Both forms outperformed the baseline for producing documents linked to topics with strengths above 0.7 ($p<=.05$), but for strength of associations on or above 0.9, the nouns only variant produced a significant increase in the number of documents, while nouns with adjectives did not. This suggests there are some advantages to reducing the data set to only nouns.

None of the POS weighted trials provided statistically significant gains in word intrusion detection compared to the trials where the news articles were simply lemmatised. The implication is that lemmatisation may be a primary contributor to improvements in word intruder detection in these trials, but further testing is required to assess this. Whether the lemmatisation played a role in the poor performance on the Wall Street Journal corpus also needs to be evaluated.

The POS weighting did not appear to reduce the variability in the top ten words generated in multiple modelling runs for the same (restructured) data set, for the same *number of topics*. Between run variability was observed both during manual reviews of the topics generated, and in terms of the variance of the observed coherence scores, of word intrusion detection, in the strength of association between a document and its primary topic, and in the number of descriptors linked to a topic. This variability is of concern if the end-user runs the topic model only once and draws conclusions about the corpus or individual documents based on that single run.

One conclusion suggested from the POS weighting trials is that reducing large datasets to nouns, or nouns and adjectives, can achieve improved run-time performance for topic modelling with no loss of topic coherence. Lemmatisation appears important to enhancing topic coherence, but based on the trials conducted in this study, it is not possible to state whether POS weighting alone, without lemmatisation, would be useful. These trials also suggest topic coherence may benefit from reducing the dataset to nouns, with or without adjectives, when the number of topics generated is large.

6.5 Weight by Named Entities

The Named Entity (NE) trials were perhaps the trials of most interest, and the trials that were the most disappointing. By focusing on a much narrower set of named entity themes, it was hoped that the topics would meaningfully link the people, locations, events and organisations in the source documents. This is particularly appealing for applied fields of research such as the digital humanities. However, the topics generated in the NE weighted trials sometimes produced misleading associations as topics contained word sets from multiple, unconnected themes in their top twenty most frequent words (e.g. *geologicalsurvey* and *StarTrek* in a single topic, as shown in the Supplementary Table A.10). This was particularly noticeable in the trials that reduced articles to only NE. The presence of misleading associations undermines the usefulness of directly reviewing and interpreting topics, as the reader must drill-down to any documents associated with

the topic to determine if the co-occurrence of particular entities reveals a real connection or is just an artefact of the topic modelling algorithm.

In addition to potentially misleading associations, the NE trials showed variability in the topics generated between runs of the same *number of topics* setting, for the same dataset, even for the substantially reduced dataset in the NE only trials. This meant that any associations identified in any given topic were not necessarily stable from run to run.

The approach to resolve multi-termed entities to a single identifier (e.g. *geological* and *survey* combined to *geologicalsurvey*) was too simple. This approach worked well when an entity had a consistent form of reference (e.g. always 'SanFrancisco'), but when entities were introduced then abbreviated, this simple approach was not effective. This was particularly the case for names (for example, introduced as *Peter Pan* but subsequent references are shortened to *Peter*), but this was also the case for corporations, or sporting teams, and even multi-termed nouns such as *geological survey* that were reduced in subsequent references to *survey*.

More sophisticated entity linking may yield better outcomes, but this is not an easy task. For example, Dunietz & Gillick (2014) describe a six step process to tag, extract and resolve tokens (using Freebase) to entity link a New York Times corpus. It may not be cost effective to do such involved processes in applied fields using novel data sources, such as 1800's regional newspapers (as in Templeton et al. (2011)). It is also questionable whether such entity linking is needed. Every topic generated from the original data that included *geological*, also included *survey* in the top twenty words for that topic (12 topics in total), and out of 212 topics that included *francisco*, 200 of these topics also included *san*. Given that topic modelling has been used in the past for named entity disambiguation (e.g. see Bhattacharya & Getoor (2006)) such co-occurrences are not unexpected.

Therefore, based on the qualitative review of the topics generated, the method of entity linking used in this study was not effective or perhaps even needed. Furthermore, the usefulness of NE weighting was undermined by run to run variability in the generated topics. The mixing of multiple themes in the same topic created spurious links, which makes such an approach unsuitable when the goal for topic modelling is to have a human reviewer read the top ten to twenty words in a topic to gain insights into the themes in the texts.

6.6 Evaluation Specific Considerations

In this study, the Information Retrieval (IR) evaluations indicated no differences between the POS weighting trials and the baseline. While this may be true, and certainly the other evaluations used in this study support this for some forms of document restructuring, it is also the case that Wei & Croft (2006) found optimum average precision scores were obtained when setting the number of topics to 800 for the SJMN corpus. This is far more that the maximum of 500 trialled in this study. As pointed out in Deveaud et al. (2013), if topic settings are too coarse or too fine grained, the topic many not reflect the queries in the IR tasks, undermining performance on such evaluations.

As to the number of topics, Wallach et al. (2009) suggests topic models have a natural optimum, after which adding additional topics makes the topics too fine grained, reducing topic quality. The observed coherence tests and the Word Intrusion tasks suggest that the 100 and 200 topics settings produced better results than the 20, 50 and 500 topic settings. This is less than the 800 topics found by Wei & Croft (2006) to be most effective for the IR task.

All configurations trialled showed variability in the topics produced from run to run, for the same setting of *number of topics* and using the same dataset. As an example, consider the 100 topic model for nouns, with uniform weighting across paragraphs, where the average observed coherence scores in each of the ten runs range from 0 to 0.48 (M=0.17, SD=0.08, N=1,000) and the average word intrusion detection rates range from 75% to 90% (M=0.825, SD=0.05, N=1,000). Given such variability, there is a question of whether 10 runs was enough to identify true differences between the forms of restructuring that were trialled. Additionally, all topic modelling was done using the hyper-parameter optimisation setting in Mallet. It is unclear whether the variability between runs would be reduced if a standard set of hyper parameters were used between runs. Tuning such parameters, and avoiding fine grained topic settings (e.g. 500 topics), may avoid the spurious linking of multiple themes in a single topic that was noted in section 5.7. Future studies may consider such adjustments.

6.7 Conclusion

In summary, the results of this study suggest reducing the dataset to only nouns speeds topic generation with no loss of topic semantic coherence. Here topic coherence is measured by an NPMI observed coherence measure, and on an automated word intrusion detection task, over topics generated based on the SJMN corpus. Such data reductions may even aid topic coherence if 200 or 500 topics are generated. There appeared to be no advantage to topic coherence from weighting the lead paragraph over the remaining SJMN article text, or from identifying and weighting named entities.

Topics need to be interpretable, valid and reliable to be able to provide useful insights into the source texts. That topics in some trials contained words from unrelated themes, suggests drawing inferences based solely on reading the top twenty most frequent words in a topic may not always be valid, and the topics may not always provide a single, interpretable theme. The run to run variability of the topics also raises questions about the reliability of any inferences drawn from the top ten to twenty topic words. Therefore, it would be appropriate to include other techniques apart from topic modelling in applied research, to confirm any inferences drawn about the source text. Such triangulation is an established practice in many fields, and topic modelling can be a useful tool within such a broader set of research methods.

Chapter 7

Conclusion

The research question of interest in this study was whether restructuring the documents to place greater weight on the nouns, adjectives and verbs, can produce more meaningful topics when topic modelling. The results of this study suggest the answer to this question is a provisional yes, but the effectiveness of such manipulations for the news corpus were sensitive to the number of topics generated. For larger numbers of topics (200 and 500), lemmatising the input text and reducing it to nouns only before topic modelling produced topics that were more coherent, faster. This study has demonstrated that lemmatising and reducing source texts to simply the nouns for a 200 topic model increased the coherence of topics on a normalised point-wise mutual information score (M=0.171, SD=0.081, compared to the baseline at $M = 0.162$, SD=0.087), and on an automated word intrusion detection measure (M=0.869, SD=0.028, compared to the baseline M=0.803, SD=0.032). Both increases were statistically significant (p<.05). Reducing the data set to only nouns also made the topic modelling faster, reducing run times from the baseline average M=92 mins (SD=1 min) to M=75 mins (SD=3 mins). Such speed improvements would be an advantage when analysing large data sets. Additionally, reducing source data to only nouns improved the number of documents with a strong association to their primary topic, suggesting the topics were perhaps more representative of the source articles. Such manipulation of the source text did not affect performance on an information retrieval task, nor did it produce a statistically significant change to topic coherence when the number of topics generated was 100 or less, though topic model generation was still faster, compared to the baseline.

Reducing the source data to the POS that conveyed the semantic content was in a sense a relatively rough way of pseudo entity and concept identification. True entity and concept identification is a very complex process, due to the need to resolve acronyms, synonyms, word collocations, and possibly even to recognise aggregates and components of any underlying entities of interest. This study trialled using a named entity recogniser, and retokenising multi-termed named entities to a single token with the goal to capture, for example, *United Kingdom* as a single entity *unitedkingdom*. While the goal was to generate more meaningful topics that better recognised named entities, this type of simplistic restructuring of documents did not produce more coherent topics. Unexpectedly, this approach did generate topics that produced better than baseline performance for information retrieval, and had stronger average document to topic associations than the baseline.

Future work in this area could include a focus on better entity handling, or on moving beyond Latent Dirichlet Allocation (LDA) to examine how document restructuring affects other topic modelling algorithms. While the named entity trials in this study were not successful in generating more meaningful, valid and reliable topics, the idea that a researcher can move beyond synonyms, acronyms and other terminology differences to reduce documents to a set of unique concepts or entities, and meaningfully cluster those elements, is attractive for applied research over digitised documents. Such an approach was proposed by Rajagopal et al. (2013), who suggested reducing documents to a set of unambiguous concepts and clustering these without the probabilistic inferences found in LDA. Such an approach may overcome the issues of the reliability and validity of the modelled topics identified in this study. However, the reduction of source documents to such concepts is a challenge. For entity resolution, Rajagopal et al. (2013) base their approach on DBPedia (Lehmann et al., 2014) and WordNet (Fellbaum, 1998). In their study on entity linking, Dunietz & Gillick (2014) use Freebase. In domains such as biomedical research, specialised lexicons, vocabularies and ontologies have been developed that could be harvested for more sophisticated identification of domain specific entities and concepts. Where the goal is knowledge discovery, rather than prediction, and where the desire is to find meaningful, valid and reliable associations to make inferences from the source data, an approach of more sophisticated entity linking and a non-probabilistic means of clustering may be worth considering.

The focus of this study has been to examine the role for restructuring data sets to produce more semantically coherent topic models. This has only been investigated in this study using the LDA algorithm, which as mentioned in the Background chapter provides a flat representation of topics. The POS trials suggest that the LDA algorithm is relatively robust to changes in document structure. As pointed out in Li & McCallum (2006), the use of a single Dirichlet distribution when sampling topic proportions in documents can affect the LDA algorithm's ability to handle more fine grained topics when certain topics co-occur more frequently than others. It is possible that this is the case in the SJMN news corpus, and it may be worth experimenting with a more sophisticated topic modelling approach such as the Pachinko Allocation Model (PAM) algorithm. The PAM generates associations between topics in addition to the word to topic relations (Li & McCallum, 2006). For example, individual topics could have themes for particular sports, but those topics can be correlated by being sports related. An issue for the PAM approach is that it is slower to build both topics and topic correlations than to generate an LDA topic model (Li & McCallum, 2006). Of interest, is whether efforts to reduce noise in a data set using lemmatisation and selecting only content words would produce more semantically coherent topics, that form useful topic correlations, faster when using the PAM algorithm, and whether any speed gains from such noise reduction make the PAM more useful with large, complex data sets.

In conclusion, this study suggests that if the desire is to topic model a large, complex data set such as a news corpus, with the goal to generate topics that are semantically coherent, interpretable, and meaningful, then a researcher should consider reducing the data set through lemmatisation, and selection of only those POS that deliver the content of interest to the research. Such a reduction will at least improve the speed with which

the topic model is generated, but it may also produce less noisy topics that are more semantically coherent and more useful.

Appendix A

Supplementary Material

TABLE A.1: *Example document restructuring - Part I*
This shows the different forms of data restructuring applied to article: SJMN91-06081009.
(Note, the texts are all converted to lower case prior to topic modelling, as described in Chapter 4, section 4.6)

Trial
Original SUNFLOWERS THRIVE IN DRY SEASON Q Can you tell me how to grow sunflowers and save the seeds? A Sunflowers are low-water-using plants so they are a good bet this summer. Plant them from seeds spaced about five feet apart when other vegetables and flowers are planted in spring. They grow quickly to huge sizes.
Lemmatised SUNFLOWERS THRIVE IN DRY SEASON Q Can you tell me how to grow sunflower and save the seed A Sunflowers be lowwaterusing plant so they be a good bet this summer Plant them from seed space about five foot apart when other vegetable and flower be plant in spring They grow quickly to huge size
Nouns-RL-U SUNFLOWERS THRIVE DRY SEASON Q sunflower seed Sunflowers plant bet summer Plant seed foot vegetable flower spring size
Nouns+Adj-RL-U SUNFLOWERS THRIVE DRY SEASON Q sunflower seed Sunflowers lowwaterusing plant good bet summer Plant seed foot other vegetable flower spring huge size
Nouns+Adj+Verbs-RL-U SUNFLOWERS THRIVE DRY SEASON Q tell grow sunflower save seed Sunflowers be lowwaterusing plant be good bet summer Plant seed space about foot apart other vegetable flower be plant spring grow quickly huge size
Nouns-RL-F SUNFLOWERS THRIVE DRY SEASON Q Q Q Q Q sunflower sunflower sunflower sunflower sunflower seed seed seed seed seed Sunflowers plant bet summer Plant seed foot vegetable flower spring size
Nouns+Adj-RL-F SUNFLOWERS THRIVE DRY SEASON Q Q Q Q Q sunflower sunflower sunflower sunflower sunflower seed seed seed seed seed Sunflowers lowwaterusing plant good bet summer Plant seed foot other vegetable flower spring huge size
Nouns+Adj+Verbs-RL-F SUNFLOWERS THRIVE DRY SEASON Q Q Q Q Q tell tell tell tell tell grow grow grow grow grow sunflower sunflower sunflower sunflower sunflower save save save save save seed seed seed seed seed Sunflowers be lowwaterusing plant be good bet summer Plant seed space about foot apart other vegetable flower be plant spring grow quickly huge size
Double Nouns-WL-U SUNFLOWERS SUNFLOWERS THRIVE THRIVE IN DRY DRY SEASON SEASON Q Q Can you tell me how to grow sunflower sunflower and save the seed seed A Sunflowers Sunflowers be lowwaterusing plant plant so they be a good bet bet this summer summer Plant Plant them from seed seed space about five foot foot apart when other vegetable vegetable and flower flower be plant in spring spring They grow quickly to huge size size

TABLE A.2: *Example document restructuring - Part II*

Trial
Double Nouns+Adj-WL-U SUNFLOWERS SUNFLOWERS THRIVE THRIVE IN DRY DRY SEASON SEASON Q Q Can you tell me how to grow sunflower sunflower and save the seed seed A Sunflowers Sunflowers be lowwaterusing lowwaterusing plant plant so they be a good good bet bet this summer summer Plant Plant them from seed seed space about five foot foot apart when other other vegetable vegetable and flower flower be plant in spring spring They grow quickly to huge huge size size
Double Nouns+Adj+Verbs-WL-U SUNFLOWERS SUNFLOWERS THRIVE THRIVE IN DRY DRY SEASON SEASON Q Q Can you tell tell me how to grow grow sunflower sunflower and save save the seed seed A Sunflowers Sunflowers be be lowwaterusing lowwaterusing plant plant so they be be a good good bet bet this summer summer Plant Plant them from seed seed space space about about five foot foot apart apart when other other vegetable vegetable and flower flower be be plant plant in spring spring They grow grow quickly quickly to huge huge size size
Double Nouns-WL-F SUNFLOWERS SUNFLOWERS THRIVE THRIVE IN DRY DRY SEASON SEASON Q Q Q Q Q Q Can you tell me how to grow sunflower sunflower sunflower sunflower sunflower sunflower and save the seed seed seed seed seed seed A Sunflowers Sunflowers be lowwaterusing plant plant so they be a good bet bet this summer summer Plant Plant them from seed seed space about five foot foot apart when other vegetable vegetable and flower flower be plant in spring spring They grow quickly to huge size size
Double Nouns+Adj-WL-F SUNFLOWERS SUNFLOWERS THRIVE THRIVE IN DRY DRY SEASON SEASON Q Q Q Q Q Q Can you tell me how to grow sunflower sunflower sunflower sunflower sunflower sunflower and save the seed seed seed seed seed seed A Sunflowers Sunflowers be lowwaterusing lowwaterusing plant plant so they be a good good bet bet this summer summer Plant Plant them from seed seed space about five foot foot apart when other other vegetable vegetable and flower flower be plant in spring spring They grow quickly to huge huge size size
Double Nouns+Adj+Verbs-WL-F SUNFLOWERS SUNFLOWERS THRIVE THRIVE IN DRY DRY SEASON SEASON Q Q Q Q Q Q Can you tell tell tell tell tell me how to grow grow grow grow grow grow sunflower sunflower sunflower sunflower sunflower sunflower and save save save save save save the seed seed seed seed seed seed A Sunflowers Sunflowers be be lowwaterusing lowwaterusing plant plant so they be be a good good bet bet this summer summer Plant Plant them from seed seed space space about about five foot foot apart apart when other other vegetable vegetable and flower flower be be plant plant in spring spring They grow grow quickly quickly to huge huge size size
Triple Nouns-WL-U SUNFLOWERS SUNFLOWERS SUNFLOWERS THRIVE THRIVE THRIVE IN DRY DRY DRY SEASON SEASON SEASON Q Q Q Can you tell me how to grow sunflower sunflower sunflower and save the seed seed seed A Sunflowers Sunflowers Sunflowers Sunflowers Sunflowers Sunflowers be lowwaterusing plant plant plant so they be a good bet bet bet this summer summer summer Plant Plant Plant them from seed seed seed space about five foot foot foot apart when other vegetable vegetable vegetable and flower flower flower be plant in spring spring spring They grow quickly to huge size size size

TABLE A.3: *Example document restructuring - Part III*

Trial
Triple Nouns+Adj-WL-U SUNFLOWERS SUNFLOWERS SUNFLOWERS THRIVE THRIVE THRIVE IN DRY DRY DRY SEASON SEASON SEASON Q Q Q Can you tell me how to grow sunflower sunflower sunflower and save the seed seed seed A Sunflowers Sunflowers Sunflowers be lowwaterusing lowwaterusing lowwaterusing plant plant plant so they be a good good good bet bet bet this summer summer summer Plant Plant Plant them from seed seed seed space about five foot foot foot apart when other other other vegetable vegetable vegetable and flower flower flower be plant in spring spring spring They grow quickly to huge huge huge size size size
Triple Nouns+Ad+Verbsj-WL-U SUNFLOWERS SUNFLOWERS SUNFLOWERS THRIVE THRIVE THRIVE IN DRY DRY DRY SEASON SEASON SEASON Q Q Q Can you tell tell tell me how to grow grow grow sunflower sunflower sunflower and save save save the seed seed seed A Sunflowers Sunflowers Sunflowers be be be lowwaterusing lowwaterusing lowwaterusing plant plant plant so they be be be a good good good bet bet bet this summer summer summer Plant Plant Plant them from seed seed seed space space space about a bout about five foot foot foot apart apart apart when other other other vegetable vegetable vegetable and flower flower flower be be be plant plant plant in spring spring spring They grow grow grow quickly quickly quickly to huge huge huge size size size
Triple Nouns-WL-F SUNFLOWERS SUNFLOWERS SUNFLOWERS THRIVE THRIVE THRIVE IN DRY DRY DRY SEASON SEASON SEASON Q Q Q Q Q Q Q Q Q Q Can you tell me how to grow sunflower sunflower sunflower sunflower sunflower sunflower sunflower sunflower sunflower sunflower and save the seed seed seed seed seed seed seed seed seed seed seed A Sunflowers Sunflowers Sunflowers Sunflowers Sunflowers Sunflowers be lowwaterusing plant plant plant so they be a good bet bet bet this summer summer summer Plant Plant Plant them from seed seed seed space about five foot foot foot apart when other vegetable vegetable vegetable and flower flower flower be plant in spring spring spring They grow quickly to huge size size size
Triple Nouns+Adj-WL-F SUNFLOWERS SUNFLOWERS SUNFLOWERS THRIVE THRIVE THRIVE IN DRY DRY DRY SEASON SEASON SEASON Q Q Q Q Q Q Q Q Q Q Can you tell me how to grow sunflower sunflower sunflower sunflower sunflower sunflower sunflower sunflower sunflower sunflower and save the seed seed seed seed seed seed seed seed seed seed seed A Sunflowers Sunflowers Sunflowers be lowwaterusing lowwaterusing lowwaterusing plant plant plant so they be a good good good bet bet bet this summer summer summer Plant Plant Plant them from seed seed seed space about five foot foot foot apart when other other other vegetable vegetable vegetable and flower flower flower be plant in spring spring spring They grow quickly to huge huge huge size size size
Triple Nouns+Adj+Verbs-WL-F: As above, but with nouns, adjectives and verbs all tripled instead of doubled,
NE-R-U : There are no named entities in the selection of text included for article SJMN91-06081009. **NE-W-U** : Matches the original format, as there are no named entities. **Double NE-W-U** : Matches the original format, as there are no named entities.

TABLE A.4: *Mean Observed Coherence (OC) Score (NPMI) by Trial and Number of Topics*

	Number of Topics					Weighted
Trial	20	50	100	200	500	Mean
Original	0.125 (0.062)	0.147 (0.082)	0.160 (0.083)	0.162 (0.087)	0.156 (0.086)	0.157
Lemmatised	0.131 (0.056)	0.153 (0.080)	0.165 (0.085)	0.17 (0.086)	0.162 (0.084)	0.163
Nouns-RL-U	0.131 (0.057)	0.157 (0.080)	0.168 (0.079)	0.171 (0.081)	0.165 (0.083)	0.165
Nouns+Adj-RL-U	0.132 (0.058)	0.158 (0.079)	0.169 (0.081)	0.173 (0.084)	0.168 (0.085)	0.168
Nouns+Adj+Verbs-RL-U	0.129 (0.057)	0.156 (0.083)	0.165 (0.083)	0.169 (0.087)	0.162 (0.085)	0.163
Nouns+Adj-RL-U	0.132 (0.060)	0.159 (0.077)	0.169 (0.079)	0.172 (0.081)	0.166 (0.083)	0.167
Nouns-RL-F	0.136 (0.064)	0.156 (0.074)	0.164 (0.075)	0.167 (0.080)	0.155 (0.089)	0.158
Nouns+Adj-RL-F	0.131 (0.060)	0.161 (0.082)	0.166 (0.078)	0.166 (0.081)	0.157 (0.089)	0.160
Nouns+Adj+Verbs-RL-F	0.131 (0.055)	0.158 (0.080)	0.165 (0.077)	0.165 (0.081)	0.154 (0.090)	0.157
Double Nouns-WL-U	0.130 (0.058)	0.157 (0.075)	0.169 (0.081)	0.173 (0.081)	0.163 (0.087)	0.165
Double Nouns+Adj-WL-U	0.133 (0.056)	0.158 (0.076)	0.17 (0.081)	0.174 (0.082)	0.164 (0.087)	0.166
Double Nouns+Adj+Verbs-WL-U	0.125 (0.055)	0.152 (0.080)	0.168 (0.080)	0.169 (0.083)	0.158 (0.087)	0.161
Double Nouns-WL-F	0.133 (0.058)	0.158 (0.079)	0.169 (0.079)	0.169 (0.081)	0.157 (0.090)	0.161
Double Nouns+Adj-WL-F	0.131 (0.059)	0.159 (0.080)	0.170 (0.078)	0.171 (0.081)	0.159 (0.091)	0.162
Double Nouns+Adj+Verbs-WL-F	0.135 (0.060)	0.154 (0.079)	0.167 (0.079)	0.166 (0.080)	0.155 (0.090)	0.158
Triple Nouns-WL-U	0.133 (0.057)	0.160 (0.075)	0.171 (0.078)	0.174 (0.082)	0.160 (0.091)	0.164
Triple Nouns+Adj-WL-U	0.134 (0.057)	0.159 (0.073)	0.173 (0.080)	0.173 (0.082)	0.162 (0.090)	0.165
Triple Nouns+Ad+Verbsj-WL-U	0.129 (0.056)	0.156 (0.080)	0.169 (0.080)	0.169 (0.082)	0.157 (0.088)	0.160
Triple Nouns-WL-F	0.130 (0.060)	0.158 (0.077)	0.169 (0.077)	0.166 (0.082)	0.150 (0.095)	0.156
Triple Nouns+Adj-WL-F	0.130 (0.063)	0.161 (0.078)	0.168 (0.075)	0.166 (0.081)	0.151 (0.096)	0.156
Triple Nouns+Adj+Verbs-WL-F	0.135 (0.062)	0.158 (0.079)	0.166 (0.077)	0.163 (0.081)	0.147 (0.093)	0.153

Note: Standard deviation is shown in parentheses after the mean score

TABLE A.5: *Mean Proportion of Successful Intruder Word Detections, by Trial and Number of Topics*

	Number of Topics					Weighted
Trial	20	50	100	200	500	Mean
Original	0.735 (0.067)	0.810 (0.032)	0.833 (0.043)	0.803 (0.032)	0.769 (0.016)	0.786
Lemmatised	0.805 (0.06)	0.870 (0.058)	0.893 (0.026)	0.879 (0.022)	0.840 (0.012)	0.856
Nouns-RL-U	0.885 (0.058)	0.856 (0.043)	0.872 (0.027)	0.869 (0.028)	0.808 (0.024)	0.834
Nouns+Adj-RL-U	0.840 (0.094)	0.886 (0.040)	0.879 (0.030)	0.870 (0.019)	0.820 (0.024)	0.843
Nouns+Adj+Verbs-RL-U	0.840 (0.061)	0.870 (0.038)	0.885 (0.033)	0.880 (0.025)	0.830 (0.018)	0.850
Nouns-RL-F	0.785 (0.100)	0.848 (0.033)	0.861 (0.035)	0.844 (0.021)	0.763 (0.018)	0.798
Nouns+Adj-RL-F	0.810 (0.113)	0.858 (0.044)	0.881 (0.025)	0.871 (0.028)	0.780 (0.012)	0.818
Nouns+Adj+Verbs-RL-F	0.795 (0.076)	0.874 (0.057)	0.887 (0.037)	0.868 (0.027)	0.789 (0.016)	0.823
Double Nouns-WL-U	0.845 (0.093)	0.862 (0.039)	0.894 (0.028)	0.868 (0.027)	0.810 (0.015)	0.837
Double Nouns-WL-U	0.795 (0.060)	0.894 (0.031)	0.888 (0.031)	0.882 (0.031)	0.806 (0.017)	0.838
Double Nouns+Adj+Verbs-WL-U	0.79 (0.097)	0.868 (0.041)	0.892 (0.019)	0.882 (0.016)	0.823 (0.015)	0.846
Double Nouns-WL-F	0.800 (0.082)	0.854 (0.028)	0.871 (0.023)	0.864 (0.022)	0.775 (0.025)	0.812
Double Nouns+Adj-WL-F	0.785 (0.075)	0.854 (0.048)	0.895 (0.018)	0.871 (0.028)	0.780 (0.016)	0.819
Double Nouns+Adj+Verbs-WL-F	0.775 (0.049)	0.884 (0.039)	0.896 (0.016)	0.859 (0.029)	0.794 (0.016)	0.825
Triple Nouns-WL-U	0.800 (0.103)	0.854 (0.063)	0.882 (0.031)	0.859 (0.044)	0.780 (0.019)	0.815
Triple Nouns+Adj-WL-U	0.830 (0.089)	0.870 (0.050)	0.877 (0.032)	0.862 (0.024)	0.793 (0.011)	0.824
Triple Nouns+Ad+Verbsj-WL-U	0.785 (0.082)	0.856 (0.030)	0.875 (0.021)	0.877 (0.020)	0.804 (0.020)	0.831
Triple Nouns-WL-F	0.845 (0.072)	0.876 (0.040)	0.862 (0.023)	0.827 (0.021)	0.746 (0.014)	0.788
Triple Nouns+Adj-WL-F	0.806 (0.081)	0.848 (0.061)	0.885 (0.031)	0.855 (0.024)	0.745 (0.021)	0.794
Triple Nouns+Adj+Verbs-WL-F	0.820 (0.054)	0.858 (0.046)	0.904 (0.020)	0.855 (0.021)	0.767 (0.022)	0.809

Note: Standard deviations are shown in parentheses after the mean proportion of successful detections

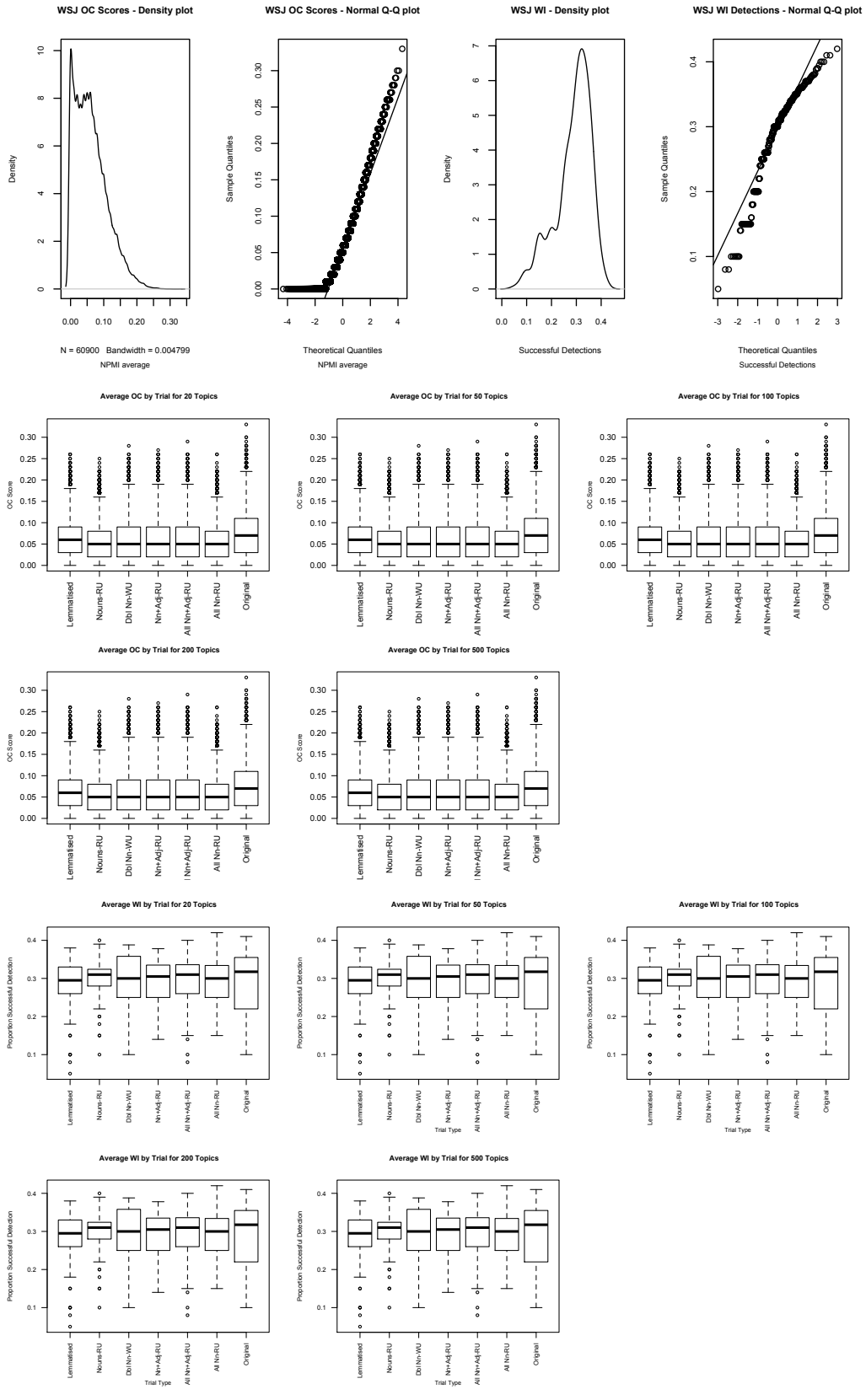


FIGURE A.1: Observed Coherence (OC) scores generated against the Wall Street Journal 1991 reference corpus
Scores are grouped by number of topics

TABLE A.6: *Topic Modelling Run Times*

	Topic Modelling time (mins)	
Trial	Mean	SD
NE-R-U	17	1
Nouns-RL-U	75	3
NE-W-U	84	2
Nouns+Adj-RL-U	86	4
NE-2W-U	92	2
Original	92	1
Lemmatised	104	2
Nouns+Adj+Verbs-RL-U	110	21
Double Nouns-WL-U	149	7
Double Nouns+Adj-WL-U	156	4
Double Nouns-WL-F	172	9
Double Nouns+Adj+Verbs-WL-U	175	4
Triple Nouns-WL-U	191	3
Double Nouns+Adj-WL-F	193	25
Triple Nouns+Adj-WL-U	212	18
Double Nouns+Adj+Verbs-WL-F	221	18
Triple Nouns-WL-F	229	15
Triple Nouns+Adj+Verbs-WL-U	246	2
Triple Nouns+Adj-WL-F	249	9
Triple Nouns+Adj+Verbs-WL-F	287	6

TABLE A.7: *Part-of-Speech Tagging Duration*

		Number of		Time to	Average Time	
	Files*	Articles	Words	POS Tag	Per Article	Per Word
Set 1	0:1-11	29,911	11,800,221	7h 40m 24s	0h 0m 0s 924ms	0h 0m 0s 2ms
Set 2	1:1-12	30,488	11,849,323	7h 58m 16s	0h 0m 0s 941ms	0h 0m 0s 2ms
Set 3	2:1-11	29,428	11,993,697	7h 22m 32s	0h 0m 0s 902ms	0h 0m 0s 2ms
Set 3	3:1	430	184,301	6m 41s	0m 0s 933ms	0m 0s 2ms
Total	90,257	35,827,542			0h 0s 925ms	0h 0s 2ms

*Files tagged in groups of no more than 2501, to avoid memory errors with the POS tagger. The timings are aggregated across sets of these individual files for easy of representation.

TABLE A.8: *Named Entity Tagging Duration*

Document	Number of	Number of		Average Time	
Set	Articles	words	Total Time	per Document	per Word
1	2501	911,092	1m 44s	0m 0s 42ms	0m 0s 0ms
2	2501	987,959	1m 58s	0m 0s 47ms	0m 0s 0ms
3	2501	982,025	1m 43s	0m 0s 41ms	0m 0s 0ms
4	2501	969,468	1m 57s	0m 0s 47ms	0m 0s 0ms
5	2501	1,010,223	1m 59s	0m 0s 48ms	0m 0s 0ms
6	2501	987,959	1m 59s	0m 0s 48ms	0m 0s 0ms
7	2501	982,025	1m 46s	0m 0s 42ms	0m 0s 0ms
8	2501	969,468	2m 4s	0m 0s 50ms	0m 0s 0ms
9	2501	1,010,223	1m 47s	0m 0s 43ms	0m 0s 0ms
10	430	184,301	22s	0s 51ms	0s 0ms
11	2400	991,497	1m 59s	0m 0s 50ms	0m 0s 0ms
12	25010	9,844,764	1h 38m 38s	0h 0m 0s 237ms	0h 0m 0s 1ms
			Average	0h 0m 0s 62ms	0h 0m 0s 0ms

TABLE A.9: *Average Number of Documents by Document-Primary Topic Association Strength*

	Document to Topic Association [0-1]						
	0.6-.69)	0.5-.59	0.4-.49	0.3-.39	0.2-.29	0.1-.19	0-0.09
Trial	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Original	4,944 (224)	7,416 (144)	12,595 (285)	19,544 (200)	25,355 (287)	13,915 (274)	193 (9)
Lemmatised	4,675 (205)	7,329 (239)	12,455 (267)	19,776 (119)	25,787 (214)	13,960 (169)	187 (14)
Nouns-RU	6,300 (265) *	9,492 (263) *	15,190 (293) *	20,946 (215) *	20,875 (325) *	8,363 (162) *	68 (7) *
Noun+Adj-R-U	5,755 (251) *	9,131 (134) *	14,115 (238) *	20,700 (373) *	22,529 (449) *	10,038 (273) *	107 (16) *
Noun+Adj+V-R-U	4,647 (171)	7,361 (265)	12,641 (229)	19,782 (369)	25,600 (304)	13,805 (249)	195 (20)
Dbl Noun-W-U	4,267 (209) *	7,006 (287) *	11,387 (283) *	18,360 (428) *	25,868 (389) *	16,630 (504) *	557 (34) *
Dbl Noun+Adj-W-U	4,310 (176) *	7,235 (173)	11,843 (219) *	18,844 (352) *	25,447 (320)	15,739 (498) *	471 (42) *
Dbl Noun+Adj+V-W-U	4,050 (170) *	6,661 (235) *	10,942 (332) *	18,217 (301) *	25,864 (367) *	17,702 (262) *	611 (29) *
Trpl Noun-W-U	4,018 (193) *	6,766 (316) *	11,040 (491) *	18,007 (359) *	26,097 (219) *	17,710 (934) *	789 (86) *
Trpl Noun+Adj-W-U	4,115 (167) *	6,975 (286) *	11,440 (250) *	18,130 (291) *	25,877 (175) *	16,885 (254) *	725 (39) *
NE-R-U	12,382 (98) *	14,585 (88) *	11,450 (132) *	13,131 (138) *	3,500 (47) *	238 (10) *	0 (0) *
NE-W-U	4,537 (166) *	7,130 (217)	12,562 (360)	21,023 (573) *	25,504 (488)	12,827 (283) *	181 (15)
Dbl NE-W-U	5,693 (102) *	9,052 (268) *	14,115 (254) *	21,758 (220) *	22,915 (195) *	8,754 (238) *	75 (11) *

* Significantly different from the baseline (i.e. the original document structure)

Note: Trials compared using Tukey all-pairwise multiple comparisons with a 95% confidence interval, calculated for each individual bin count

Counts by trial: Count across the 200 topics were aggregated for each of the 10 runs (run *a-j*) for each trial, producing 10 scores per trial (model). The means and standard deviations above were calculated across those 10 runs, for each respective trial.

The document to primary topic strength is taken from the Mallet document-topic composition file (specified in the –output-doc-topics parameter when training the topic model)

Counts for bins 0.7-.79, 0.8-.89 and 0.9+ are in the main document

TABLE A.10: *Example topics*

ID	Trial	Topic 1st 10 words	2nd 9 words
All topics from the Nouns only-100 topics trial that mention 'wine'			
1.	Nouns-RL-U -100a	wine winery valley fruit beer bottle cabernet chardonnay sauvignon flavor	napa alcohol vineyard grape sonoma california vineyards blanc soph
2.	Nouns-RL-U -100b	water wine plant tree garden flower fruit valley drought soil	winery california leaf seed gallon year rose cabernet chardonnay
3.	Nouns-RL-U -100c	check wine winery valley fruit bottle beer chardonnay cabernet	napa california sauvignon flavor vineyard grape alcohol sonoma vineyards drink
4.	Nouns-RL-U -100d	wine winery valley fruit bottle beer chardonnay cabernet napa sauvignon	flavor california vineyard grape alcohol sonoma vineyards blanc tasting
5.	Nouns-RL-U -100e	palo alto wine mountain view valley altos winery park east	menlo los california fruit cabernet chardonnay napa sauvignon vineyard
6.	Nouns-RL-U -100f	wine croatia yugoslavia army winery republic yugoslav slovenia serbia valley	fruit cabernet chardonnay sauvignon serbs napa vineyard flavor grape
7.	Nouns-RL-U -100g	check wine winery valley fruit california chardonnay cabernet bottle sauvignon	napa flavor vineyard grape sonoma vineyards tasting blanc pinot
8.	Nouns-RL-U -100h	wine winery valley fruit bottle beer chardonnay cabernet alcohol flavor	california napa sauvignon vineyard grape sonoma vineyards drink blanc
9.	Nouns-RL-U -100i	wine winery valley fruit beer bottle chardonnay cabernet napa sauvignon	flavor california vineyard grape sonoma vineyards alcohol blanc tasting
10.	Nouns-RL-U -100j	wine plant tree garden fruit flower winery soil valley california	leaf seed rose cabernet year chardonnay variety flavor sauvignon
All topics from the Named Entity only-200 topics trial that mention 'geological survey'			
11.	NE-R-U -200b	taylor richter geologicalsurvey lomaprieta pasadena hilton nichols bayarea elizabethtaylor	usgs burton liz schrader enquirer sanfrancisco menlopark sainteclaire sanandreasfault sierramadre
12.	NE-R-U-200c	bayarea santaclaracounty lomaprieta sanfrancisco richter geologicalsurvey wong caltrans coles gordon	berman usgs marin silverado santaclara sanandreas alameda mtc metropolitantransportationcommission
13.	NE-R-U-200d	alaska exxon garcia watsonville anchorage hickel greengiant exxonvaldez arco richter	lomaprieta alaskan geologicalsurvey exxoncorp broderick texaco fairbanks usgs forestservice
14.	NE-R-U-200d	richter redcross lomaprieta geologicalsurvey bayarea coles usgs wong wayne lawrence	johnwayne menlopark americanredcross median sanandreasfault taos pasadena sanandreas stanfordstadium
15.	NE-R-U-200e	taylor richter christo pasadena geologicalsurvey elliott lomaprieta usgs elizabethtaylor mathews	liz menlopark bayarea sierramadre sanandreas sanandreasfault earhart webber california
16.	NE-R-U-200e	trump george webb donaldtrump richter harmon lomaprieta bakker geologicalsurvey galileo	donald coles usgs missamerica ivana sudafed busch bozo bayarea
17.	NE-R-U-200f	bart caltrans bayarea sanfrancisco lomaprieta santaclaracounty fremont geologicalsurvey caltrain oakland	sanjose actransit bianco richter glenn californiadepartmentoftransportation coles usgs warmsprings
18.	NE-R-U-200f	ross mr jensen bayarea startrek ms hud harmon geologicalsurvey kirk	zanger mrs coles richter stern usgs mercurynews spock mccoy
19.	NE-R-U-200g	earth lomaprieta richter kahn geologicalsurvey bayarea buckley pasadena usgs fraser	coles sdepartment smithsonian menlopark feshbachs californiainstituteoftechnology fema kenkahn mountst
20.	NE-R-U-200h	salomon conner coastguard treasury richter buffett wallstreet geologicalsurvey lomaprieta salomonbrothers	homefed usgs coles gutfreund bayarea mozer salomonbrothersinc sanandreas goodrich
21.	NE-R-U-200i	ross lomaprieta richter geologicalsurvey barnes bakker coles klein bayarea usgs	wong edison preston menlopark sanandreas hampton tyler sanandreasfault potter
22.	NE-R-U-200j	intel amd faa usair richter boeing federalaviationadministration skywest geologicalsurvey intelcorp	advancedmicrodevicesinc shaw nationaltransportationsafetyboard burnett usgs lomaprieta losangeles bressler ntsb
23.	NE-R-U-200k	redwoodcity sanmateo menlopark sanmateocounty parker hollister rogers woodside paloalto sancarlos	usgs eastpaloalto bayarea geologicalsurvey lomaprieta sanbenitocounty sanfrancisco atherton santacruz

References

Aletras, N. & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, (pp. 13–22)., Potsdam, Germany. Association for Computational Linguistics. 14, 24

Bhattacharya, I. & Getoor, L. (2006). A latent dirichlet model for unsupervised entity resolution. In *Proceedings of the Sixth SIAM International Conference on Data Mining*, (pp. 47–58)., Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. 15, 48

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. 2, 10, 11

Blei, D. M., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested chinese restaurant process. In Thrun, S., Saul, L., & Schölkopf, S. (Eds.), *Advances in Neural Information Processing Systems 16*, (pp. 17–24)., Cambridge, MA, USA. MIT Press. 11

Blei, D. M. & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (pp. 113–120). 10, 11

Blei, D. M. & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35. 11

Blei, D. M., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. 1, 10, 23

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In Chiarcos, C., de Castilho, R. E., & Stede, M. (Eds.), *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, (pp. 31–40)., Tübingen, Germany. Narr Verlag. 14, 24

Boyd-Graber, J. L. & Blei, D. M. (2009). Syntactic topic models. In Koller, D., Schuurmans, D., Bengio, Y., & Bottou, L. (Eds.), *Advances in Neural Information Processing Systems*, (pp. 185–192). Curran Associates, Inc. 13

Brett, M. R. (2012). Topic modeling: A basic introduction. *Journal of Digital Humanities*, *2*(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/. 2

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., & Culotta, A. (Eds.), *Advances in Neural Information Processing Systems 22*, (pp. 288–296). Curran Associates, Inc. 13

Chen, H., Branavan, S. R. K., Barzilay, R., & Karger, D. R. (2009). Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, (pp. 371–379)., Stroudsburg, PA, USA. Association for Computational Linguistics. 15

Claeskens, G. & Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 23

Darling, W. M., Paul, M. J., & Song, F. (2012). Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, (pp. 1–9)., Stroudsburg, PA, USA. Association for Computational Linguistics. 12, 13

Deveaud, R., SanJuan, E., & Bellot, P. (2013). Are semantically coherent topic models useful for ad hoc information retrieval? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (pp. 148–152)., Sofia, Bulgaria. Association for Computational Linguistics. 48

Du, L., Buntine, W., & Jin, H. (2012). Modelling sequential text with an adaptive topic model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, (pp. 535–545)., Stroudsburg, PA, USA. Association for Computational Linguistics. 15

Dunietz, J. & Gillick, D. (2014). A new entity salience task with millions of training examples. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, (pp. 205–209)., Gothenburg, Sweden. Association for Computational Linguistics. 47, 52

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Language, speech, and communication. MIT Press. 52

Finkel, J. R. (2007). Named Entity Recognition and the Stanford NER Software. `http://nlp.stanford.edu/software/jenny-ner-2007.pdf`. 8

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (pp. 363–370). Association for Computational Linguistics. 8, 21

Foster, J. J., Barkus, E., & Yavorsky, C. (2006). *Understanding and Using Advanced Statistics: A Practical Guide for Students*. SAGE Publications. 26

Francis, W. N. & Kucera, H. (1979). Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US. 7

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Saul, L., Weiss, Y., & Bottou, L. (Eds.), *Advances in Neural Information Processing Systems 17*, (pp. 537–544). MIT Press. 12, 13

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, *18*(1), 1–35. 2

Han, X. & Sun, L. (2012). An entity-topic model for entity linking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, (pp. 105–115)., Stroudsburg, PA, USA. Association for Computational Linguistics. 15

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, (pp. 50–57)., New York, NY, USA. ACM. 10

Jiang, J. (2009). Modeling syntactic structures of topics with a nested HMM-LDA. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, (pp. 824–829)., Washington, DC, USA. IEEE Computer Society. 12, 13

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition* (2nd Edition ed.). Prentice Hall Series in Artificial Intelligence. Upper Saddle River, N.J.: Pearson Prentice Hall. 8

Kim, H., Sun, Y., Hockenmaier, J., & Han, J. (2012). ETM: Entity topic models for mining documents associated with entities. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, (pp. 349–358)., Washington, DC, USA. IEEE Computer Society. 15

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), Gothenburg, Sweden.* 3, 4, 14, 16, 17, 20, 24, 28, 44

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2014). DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*. 52

Li, W. & McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (pp. 577–584)., New York, NY, USA. ACM. 11, 53

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational*

Linguistics and Intelligent Text Processing - Volume Part I, CICLing'11, (pp. 171–189)., Berlin, Heidelberg. Springer-Verlag. 7, 8

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. 7

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu. 23

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *EMNLP*. 2, 3, 14, 16, 24, 41

Nelson, R. K. (2010). Mining the dispatch. http://dsl.richmond.edu/dispatch/pages/intro. 2

Newman, D., Chemudugunta, C., & Smyth, P. (2006). Statistical entity-topic models. In *Knowledge Discovery and Data Mining*. 15

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, (pp. 100–108)., Stroudsburg, PA, USA. Association for Computational Linguistics. 14, 16, 24

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. 6

Porter, M. (1997). An algorithm for suffix stripping. In K. Sparck Jones & P. Willett (Eds.), *Readings in Information Retrieval* (pp. 313–316). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 6

Rajagopal, D., Olsher, D., Cambria, E., & Kwok, K. (2013). Common sense-based topic modeling. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM 2013, (pp. 6:1–6:8)., New York, NY, USA. ACM. 52

Salton, G. & McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill computer science series. McGraw-Hill International. 11

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, *2*(1). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/. 2, 3

Shu, L., Long, B., & Meng, W. (2009). A latent topic model for complete entity resolution. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, ICDE '09, (pp. 880–891)., Washington, DC, USA. IEEE Computer Society. 15

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. 10, 11

Templeton, C., Brown, T., Battacharyya, S., & Boyd-Graber, J. (2011). Mining the dispatch under supervision: Using casualty counts to guide topics from the richmond daily dispatch corpus. In *Chicago Colloquium on Digital Humanities and Computer Science*. 47

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, (pp. 173–180). Association for Computational Linguistics. 7, 19

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (pp. 1105–1112)., New York, NY, USA. ACM. 48

Wang, H., Zhang, D., & Zhai, C. (2011). Structural topic model for latent topical structure analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, (pp. 1526–1535)., Stroudsburg, PA, USA. Association for Computational Linguistics. 15

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, (pp. 697–702). 26

Webber, B., Egg, M., & Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, *18*, 437–490. 9

Wei, X. & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, (pp. 178–185)., New York, NY, USA. ACM. 23, 26, 48

Yang, T.-I., Torget, A. J., & Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, (pp. 96–104)., Stroudsburg, PA, USA. Association for Computational Linguistics. 15

Zhu, X., Blei, D. M., & Lafferty, J. (2006). TagLDA: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, Madison. 15