A Study of the Use of Keyword and Keyphrase Extraction Techniques for Answering Biomedical Questions

By

Jiwei Guan

A thesis submitted to Macquarie University
for the degree of Master of Research
Department of Computing
January 2016

MACQUARIE
University
SYDNEY·AUSTRALIA

Typeset in $\LaTeX\,2_\varepsilon$.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Jiwei Guan

Acknowledgements

I would like to thank my supervisor, Dr Diego Mollá Aliod for his advice, patience and motivation. His guidance helped me throughout the MRes research project and the writing of this thesis. I also thank Adjunct Professor Hanna Suominen, a researcher at National Information and Communications Technology Australia in Canberra, for her advice.

Abstract

This research project explored the use of automatic keyword and keyphrase extraction techniques as a means to generate answers to biomedical questions. Keywords and keyphrases provide an essential way to present the topic of a given document and can help readers access core information in text. Keyword and keyphrase extraction techniques are typically used in information retrieval tasks. The purpose of this project is to select the suitability of these techniques in order to extract key concepts in training dataset of BioASQ shared task, as a first step towards achieving query-based abstractive summarisation. The outputs are measured by F1-metric to distinguish the performance of each technique.

Contents

List of Figures

List of Tables

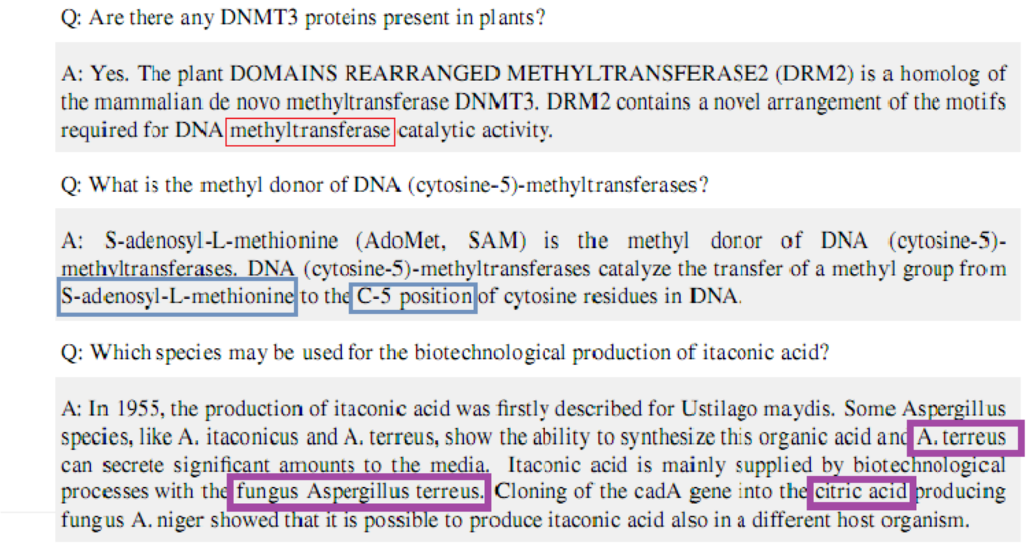
1 Introduction

1.1 Introduction

With the information explosion we face nowadays, and large storage space available at reduced costs, scientific information is expanding and the number of scientific articles published is increasing. Scientific articles are recorded in text. And text is unstructured data within digital forms, which may contain underlying information. This is particularly the case of medical domain. More medical information is available in text than ever before and medical research publications retrieval has been attracting both computing and medical research communities, especially in the biomedical domain. For instance, the National Center for Biotechnology Information (NCBI http://www.ncbi.nlm.nih.gov/) in America is collecting and maintaining biomedical literature to conduct biomedical research, and is computerising biomedical information. NCBI is a sub program in National Library of Medicine (NLM https://www.nlm.nih.gov). With such an overwhelming amount of biomedical literature, there is much interest in techniques to explore underlying information and uncover new knowledge.

Medical practitioners and clinicians have clinical questions which need to be answered to diagnose patients. To make best decisions, practitioners synthesise all of the important information about the patient, relevant research, and experience with previous patients to determine the best course of action (Eddy, 2005). During the process, practitioners review relevant literature and use search engines to find evidence to apply in their medical practice. However, it is still extremely difficult for physicians to obtain the most useful evidence in a large literature. For example, when a practitioner faces a patient, the practitioner uses a fundamentally human process called the "art of medicine" and "clinical judgment" which are empirical. This process may not support the practitioner to make right decisions, because such approaches are highly subjective rather than depending on facts or evidence. For better diagnosis, there is a way that is needed to review and find the evidence. Evidence-based medicine (EBM) attempts to educate practitioners using medical literature applied to clinical practice.

Exploring evidence in huge biomedical text is still a challenging task. This is motivated by the desire of many important clinical practices which are quicker, more convenient, more consistent and more efficient than current practice. Hidden evidence in text can be terms including keywords and keyphrases. Terms can help researchers, biologists and physicians to summarise a large body of literature, make use of textual knowledge and support health-care decision-making. Since manually searching terms is time-consuming and expensive, automated term extraction techniques can save time and economy. Therefore, automated extraction and automated knowledge acquisition are necessary and important. This research project tries to extract keywords and keyphrases that could be evidence applied to medical practice from the BioASQ dataset.



Q: Are there any DNMT3 proteins present in plants?

A: Yes. The plant DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2) is a homolog of the mammalian de novo methyltransferase DNMT3. DRM2 contains a novel arrangement of the motifs required for DNA methyltransferase catalytic activity.

Q: What is the methyl donor of DNA (cytosine-5)-methyltransferases?

A: S-adenosyl-L-methionine (AdoMet, SAM) is the methyl donor of DNA (cytosine-5)-methyltransferases. DNA (cytosine-5)-methyltransferases catalyze the transfer of a methyl group from S-adenosyl-L-methionine to the C-5 position of cytosine residues in DNA.

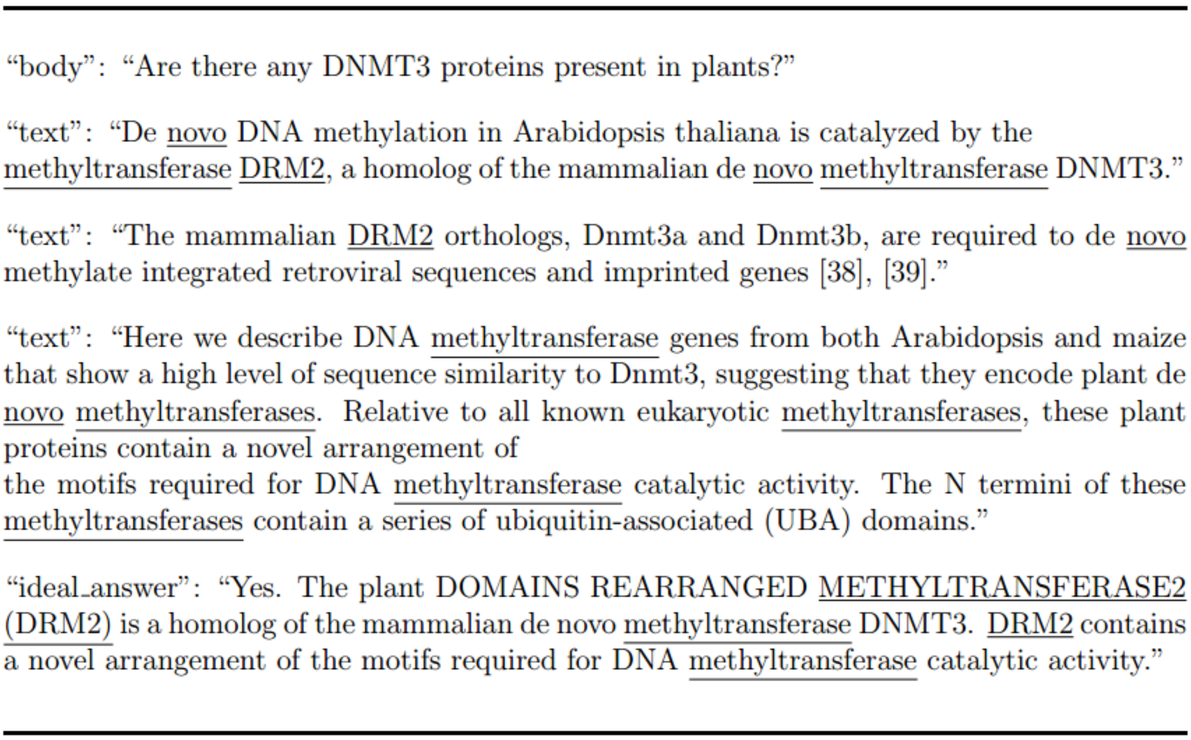
Q: Which species may be used for the biotechnological production of itaconic acid?

A: In 1955, the production of itaconic acid was firstly described for Ustilago maydis. Some Aspergillus species, like A. itaconicus and A. terreus, show the ability to synthesize this organic acid and A. terreus can secrete significant amounts to the media. Itaconic acid is mainly supplied by biotechnological processes with the fungus Aspergillus terreus. Cloning of the cadA gene into the citric acid producing fungus A. niger showed that it is possible to produce itaconic acid also in a different host organism.

FIGURE 1.1: Questions and ideal answers in BioASQ final report
Source: Paliouras and Krithara (2015)

BioASQ (http://www.bioasq.org/) is an ongoing European project and is a large project to search and summarise biomedical text. BioASQ has two tasks: task A is Online Biomedical Semantic Indexing and task B is Biomedical Semantic Question Answering which requires participants to accept syntactically well-formed and often quite complex English questions (Paliouras & Krithara, 2015). Figure 1.1 shows the examples of biomedical questions and ideal answers in BioASQ final report (Paliouras & Krithara, 2015). Potential keywords and keyphrases have been highlighted in red, blue and purple. Figure 1.2 shows a part of the

BioASQ dataset. The question is the "body" and relevant passages are the "text". Passages are relevant answers in English and are text fragment from literature. These biomedical questions and "ideal_answer" are written by biomedical experts with social networks, which reflect the real-life information needs. The content of a "ideal_answer" is the standard and used in evaluation. Terms are highlighted in the "text" and the "ideal_answer" in Figure 1.2. The common terms between "text" and "ideal_answer" can be selected as answers for biomedical questions.



"body": "Are there any DNMT3 proteins present in plants?"

"text": "De novo DNA methylation in Arabidopsis thaliana is catalyzed by the methyltransferase DRM2, a homolog of the mammalian de novo methyltransferase DNMT3."

"text": "The mammalian DRM2 orthologs, Dnmt3a and Dnmt3b, are required to de novo methylate integrated retroviral sequences and imprinted genes [38], [39]."

"text": "Here we describe DNA methyltransferase genes from both Arabidopsis and maize that show a high level of sequence similarity to Dnmt3, suggesting that they encode plant de novo methyltransferases. Relative to all known eukaryotic methyltransferases, these plant proteins contain a novel arrangement of the motifs required for DNA methyltransferase catalytic activity. The N termini of these methyltransferases contain a series of ubiquitin-associated (UBA) domains."

"ideal_answer": "Yes. The plant DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2) is a homolog of the mammalian de novo methyltransferase DNMT3. DRM2 contains a novel arrangement of the motifs required for DNA methyltransferase catalytic activity."

Figure 1.2: Real examples of question, relevant answers and "ideal_answer" in BioASQ dataset

Currently, machines cannot independently produce an "ideal_answer" from huge literature. EBM-Group (1992) foresaw "practical approaches to making evidence-based summaries easier to apply in clinical practice, many based on computer technology, will be developed and expanded." In future, the ultimate goal of query-based summarisation is to automatically produce an "ideal answer" for a medical question. In BioASQ task B, relevant passages corresponding to questions could potentially answer biomedical questions. Therefore, the first step is whether extraction techniques can be used to find suitable terms in relevant passages to identify biomedical answers, which can move closer to the "ideal_answer".

These terms can be used in query-based summarisation and as core concepts of abstractive summarisation to answer biomedical questions, which are user-understandable. Then the purpose of this research project is to review and implement extraction approaches to explore terms from such passages. It tests extraction techniques whether they can successfully identify terms, that is, keywords and keyphrases in the "ideal_answer". In Section 1.2 we present the history of EBM and its concepts. In Section 1.3 we introduce the current biomedical database systems. Section 1.4 focuses on text summarisation, and keywords and keyphrases for the summary. In Section 1.5 we present aims of the research project. In Section 1.5, we present the methodology for the project. Finally, Section 1.7 outlines the structure of this thesis.

1.2 Background of Evidence-based medicine

Since EBM-Group (1992) published "Evidence-based medicine: A new approach to teaching the practice of medicine" in the *Journal of the American Medical Association*, it has become a new subject in medical science. The term EBM was first used by David Sackett and colleagues at McMaster University in Ontario. EBM has developed from the early 1990s to search clinical literature and applications of formal rules for evidence. Later, the definition of EBM was refined as "EBM is integrating the best research evidence with clinical expertise and patient values to achieve the best possible patient management" (Glasziou et al., 2010). The evidence is searched through different records, such as diagnosis records, clinical literature, research publications, published surveys of front-line clinicians, and written records of seminars and relevant studies. EBM intends to summarise evidence with comprehensive literature collections and it combines systematic review and external or formal evidence to apply to medical practices. However, it is noted that EBM is not a new process — "clinicians and physicians have always strived to make conscientious, explicit, and judicious use of current best available evidence in making decisions about their patients" (Sackett et al., 1996). Therefore, the application of EBM can help clinicians and researchers in decision making. Three factors are in the process of EBM: best available clinical evidence, experience of the individual clinician, and patient needs, desires and resources. In Figure 1.3, Pamela Corley & Adrian Follette (2003) concluded the interaction of EBM with the three factors.

EBM has two branches: Evidence-Based Healthcare (EBHC) and Evidence-based practice (EBP). EBHC is about the health policy and management decisions rather than the individual patient. For instance, the evidence indicates that well-designed physical settings play an important role in making hospitals safer and more healing for patients, and better places for staff to work (Ulrich et al., 2008). EBP is an approach to healthcare where health professionals use the best evidence possible and the most appropriate information available to make clinical decisions for individual patients (McKibbon, 1998). Thus EBM is also the formal process of integrating personal experience, evidence and other factors to identify clinical questions. This research project focuses on answering clinical questions using extracted terms, which belongs to EBP.

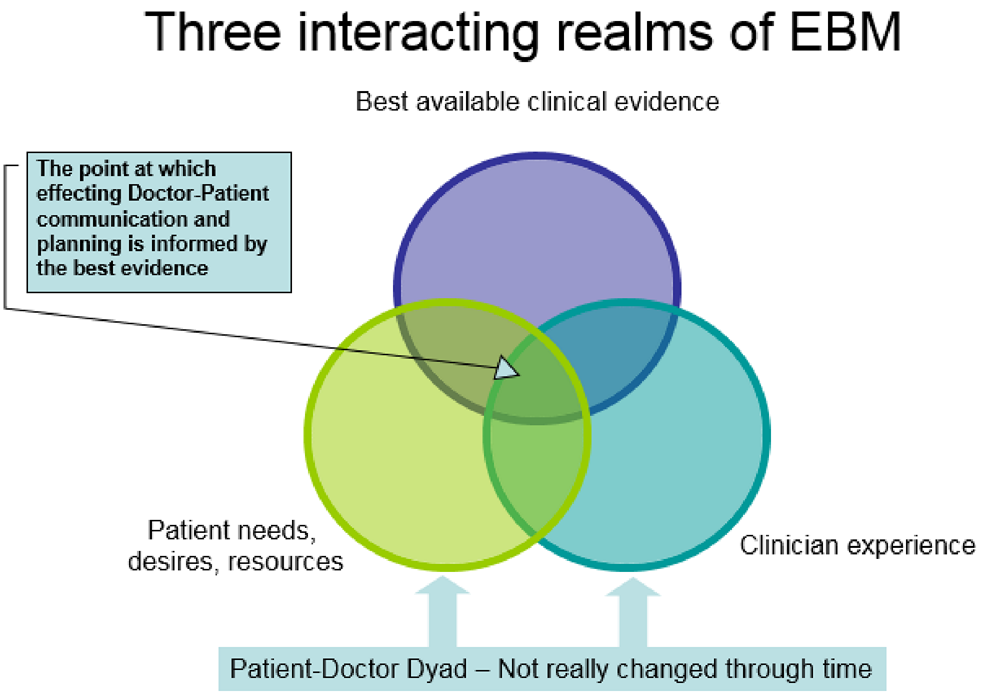

Three interacting realms of EBM
Best available clinical evidence
The point at which effecting Doctor-Patient communication and planning is informed by the best evidence
Patient needs, desires, resources
Clinician experience
Patient-Doctor Dyad – Not really changed through time

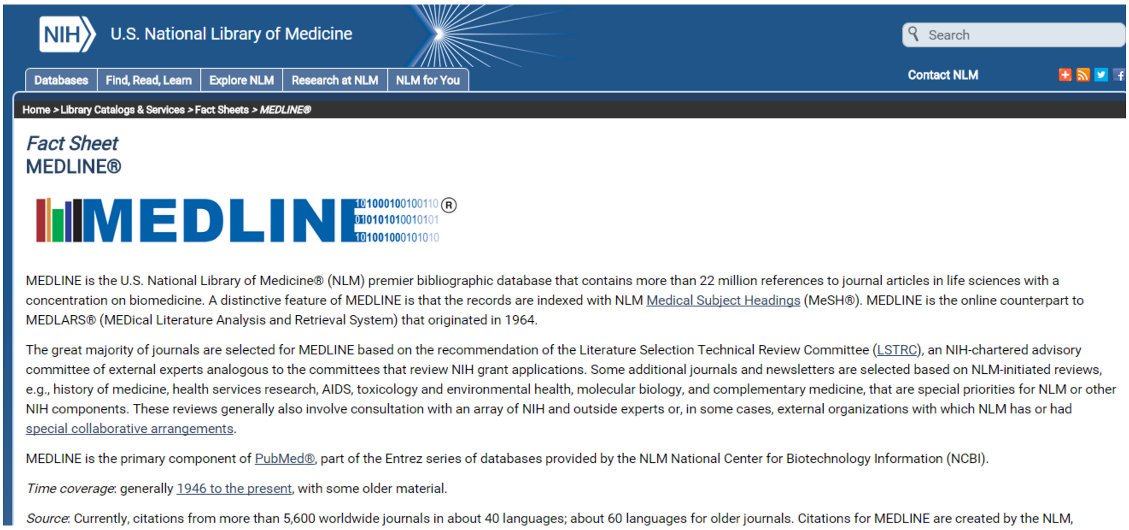
Figure 1.3: Interacting realms of EBM
Source: http://www.usc.edu/hsc/nml/portals/orientation/, Corley el al. (2003)

1.3 Biomedical database systems

As medical publications are normally composed of text with restricted access due to copyright reasons, most medical research publications and journals are not freely accessible. But there is online index named MEDLINE (http://www.ncbi.nlm.nih.gov/guide/literature) for biomedical publications which mainly contain abstracts of articles. **MEDLINE** (Medical Literature Analysis and Retrieval System Online) is the biggest database of Medical Literature Analysis and Retrieval System (MEDLARS) in NLM and contains journal citations and abstracts for biomedical literature from around the world. It has at least 15 million references to journal articles in the life sciences in 2006, and it is increasing at a rate of more than 10% each year (Ananiadou et al., 2006). MEDLINE does not usually contain full length articles. Figure 1.4 shows a description of MEDLINE. Besides MEDLINE, there are other successful biomedical information databases. They include but are not limited to:

PubMed (http://www.ncbi.nlm.nih.go/pubmed/) is a free biomedical literature search engine for MEDLINE. The site provides access to MEDLINE and links to full text articles. It contains more than 24 million citations for biomedical literature, life science journals and online books from MEDLINE. Citations may include links to full-text content from PubMed Central and publisher websites. There is a guide about MEDLINE and PubMed: https://www.nlm.nih.gov/bsd/pmresources.html. With the popularity of MEDLINE, The

Chinapubmed(http://www.chinapubmed.net/) has been developed since 2005 and text resources in Chinapubmed are bilingual: English and Chinese. Figure 1.5 is a screenshot of the entry screen of PubMed.


NIH U.S. National Library of Medicine
Search
Databases
Find, Read, Learn
Explore NLM
Research at NLM
NLM for You
Contact NLM
Home > Library Catalogs & Services > Fact Sheets > MEDLINE®
MEDLINE 101000100100110 010101010010101 101001000101010 ®

Fact Sheet
MEDLINE®

MEDLINE is the U.S. National Library of Medicine® (NLM) premier bibliographic database that contains more than 22 million references to journal articles in life sciences with a concentration on biomedicine. A distinctive feature of MEDLINE is that the records are indexed with NLM Medical Subject Headings (MeSH®). MEDLINE is the online counterpart to MEDLARS® (MEDical Literature Analysis and Retrieval System) that originated in 1964.

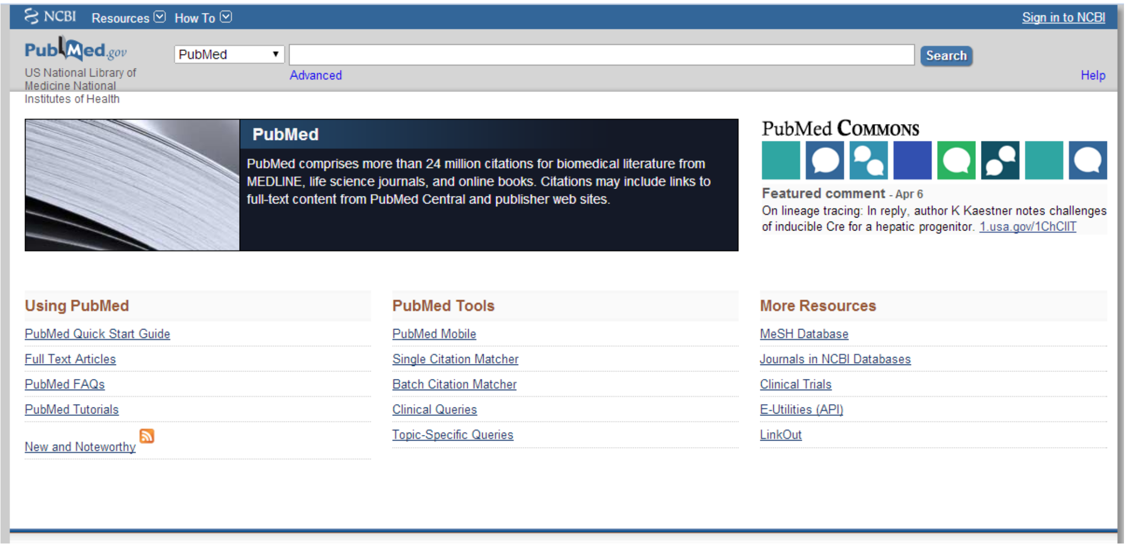
The great majority of journals are selected for MEDLINE based on the recommendation of the Literature Selection Technical Review Committee (LSTRC), an NIH-chartered advisory committee of external experts analogous to the committees that review NIH grant applications. Some additional journals and newsletters are selected based on NLM-initiated reviews, e.g., history of medicine, health services research, AIDS, toxicology and environmental health, molecular biology, and complementary medicine, that are special priorities for NLM or other NIH components. These reviews generally also involve consultation with an array of NIH and outside experts or, in some cases, external organizations with which NLM has or had special collaborative arrangements.

MEDLINE is the primary component of PubMed®, part of the Entrez series of databases provided by the NLM National Center for Biotechnology Information (NCBI).

Time coverage: generally 1946 to the present, with some older material.

Source: Currently, citations from more than 5,600 worldwide journals in about 40 languages; about 60 languages for older journals. Citations for MEDLINE are created by the NLM,

FIGURE 1.4: Screenshot of MEDLINE fact sheet
Source: https://www.nlm.nih.gov/pubs/factsheets/medline.html



PubMed

PubMed comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Featured comment - Apr 6
On lineage tracing: In reply, author K Kaestner notes challenges of inducible Cre for a hepatic progenitor. 1.usa.gov/1ChCIIT

FIGURE 1.5: Screenshot of PubMed entry
Source: http://www.ncbi.nlm.nih.gov/pubmed/

Textpresso (http://www.textpresso.org or WormBase http://www.wormbase.org) uses the ontology-based approach to extract and retrieve biological literature. It major classifies biological concepts such as gene, allele, cell or cell group, and phenotype and their related two objects such as association or regulation.

Query Chem (http://www.querychem.com/) is a web program that integrates chemical structure and text-based searching using publicly available chemical databases and Google's Web Application Program Interface (API). The advantage of Query Chem is that the user can use the database without knowing the molecular names, and retrieve the compound structure and their properties.

GoPubMed (http://www.gopubmed.org/web/gopubmed/) is a search engine tool using PubMed database. It is driven by gene ontology and Medical Subject Headings (MeSH) to refine the results of PubMed.

PubMatrix (http://pubmatrix.grc.nia.nih.gov/) compares the list of terms in PubMed. The list of terms could be gene or protein names, diseases, gene functions and authors. It also outputs the associated list of terms and functionalities (Becker et al., 2003).

PubFinder (http://www.glycosciences.de/tools/PubFinder) is a web tool designed to retrieve scientific abstracts for a specific topic. It returns a selection of articles that are most relevant to a scientific topic.

iHOP (http://www.ihop-net.org/UniPub/iHOP/) provides the visualisation network of genes and proteins by scientific literature. It can access millions of PubMed abstracts as well.

1.4 Text summarisation

1.4.1 Automatic text summarisation

Although biomedical databases and extraction tools can greatly assist practitioners and medical researchers to access and search by their needs and interests, there are still no efficient ways to summarise available evidence. Even though medical database systems are already very popular and convenient, it is still not possible for users to process various information and benefit of the clinical assumptions for a complex medical decision. With the expansion and growth of medical information, it is difficult for users to search relevant journals and professional associations, explore different language websites and read reports in different languages. Automatic text summarisation was first proposed by Luhn (1958). Automatic text summarisation techniques can help researchers and practitioners to quickly find required information and determine the main idea of a given document.

Radev et al. (2002) defined the summary: " a summary can be loosely defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the originals text(s), and usually significantly less than that." Text here is used loosely and can refer to speech, multimedia, documents, reports, publications and hypertext. There are two types of automatic text summarisations: indicative summarisation and informative summarisation. The indicative summarisation points to the main idea of the original document, to help readers decide whether or not to read further. In contrast, the informative summarisation only provides the explanation to present all the related information. Both indicative and informative summarisations are contributions to the query-based summarisation — made up of answers using related information by users' query. Query-based summarisation can be used either on a single document, focusing on the topic of a single document summarisation, or multi-document summarisations, producing multiple documents summarisation. In addition, automatic query specific multi-document summarisation is the process of filtering important relevant information from the input set of documents and presenting the concise version of documents to the user (Chowdary et al., 2010).

Automatic text summarisation can also be simply classified into abstractive and extractive. An abstractive summarisation tries to understand the core concept of the given document and produces a summary in clear natural language. Specifically, it deals with reformulating important sentences, presenting a concise representation of text and assembling them as a summary, which is customised towards users' needs or provides an overall sense of the document (Jones et al., 1999; Chowdary et al., 2010). It needs to understand the original document and organise the different expressions. The purpose of extractive summarisation is to extract text fragments from an original document such as important sentences or paragraphs. This type of summarisation thus avoids any effort at deep text understanding, and it is conceptually simple and easy to implement (Gupta & Lehal, 2010), which is more robust and more feasible approach.

The current research in informatics mainly focuses on extractive summarisation and this form of summary has been proven to be effective. However, extractive summarisation still cannot provide deep understanding of the original text and it directly extracts contents by sequence. Abstractive summarisation is more human-readable and coherent because of reorganising contents, which has not yet been developed. Therefore, abstractive summarisation is the new focus of research. So far, it has not been explicitly determined how to develop and generate abstractive summarisation. Our research supports to find key concepts of query-based abstractive summarisation using extracted terms, as well as multi-documents summarisation. Terms usually contain keywords and keyphrases. Keywords are those words which can capture the topic. Keyphrases contain several words or phrases to summarise the theme of a document. Keywords and keyphrases are important means by which to condense the source text into a summary. In this research project, keywords and keyphrases are totally called terms.

1.4.2 Keywords and keyphrases for the summarisation

Many published research articles introduced the general text summarisation by keywords and keyphrases extraction(Luhn, 1958; Edmundson, 1969; Paice, 1980). Edmundson (1969) presented four sentence scoring automatic extracting methods for summarisation: key word feature, cue word feature, title word feature and sentence location feature, and defined a text summarisation system formula:

$$\mathrm{W} = \alpha_1\mathrm{C}{+}\alpha_2\mathrm{K}{+}\alpha_3\mathrm{T}{+}\alpha_4\mathrm{L}$$

where α_1, α_2, α_3 and α_4 were the parameters (positive integers) for the four methods weights respectively.

Key word feature: The key word method compiles a key glossary for each document, ideally consisting of topic words statistically selected from the body of that document.

Cue word feature: Sentences containing phrases such as "in conclusion", "in summary","in particular", "in conclusion" and "the aim of this paper" may be good indicators of relevant information (Lloret & Palomar, 2012). Although words such as "summary", "argue", "report" and "conclusion" are not important words, they have indicative functions and they tell readers the following part is important. They are more likely to be shown in the summary.

Title word feature: Normally, titles or subtitles are those words or phrases which are assigned by the author(s). Based on this, sentences that contain these words or phrases are closer to the document topic. Sentences with this feature are more likely to conclude the paragraph or the document.

Sentence location feature: Usually the first and last sentence of the first and last paragraph of a text document are more important and have a greater chance of being included in the summary (Gupta & Lehal, 2010). These positions in a document should receive close attention.

Edmundson (1969) started to use these features to produce text summarisations. More and more researchers produced text summarisations and they proposed many novel approaches. This thesis presents our research project selects extraction approaches to generate key information of summarisation that could be used to answer biomedical questions.

1.5 Aims and research questions

Currently, it is not clear how to automatically produce a correct answer that is "ideal_answer". At this stage, we can use extraction techniques to explore evidence as the summary that could be part of "ideal_answer". The overall aim of this research project is to explore the idea of using automatic extraction approaches to extract information from relevant fragments as a means to answer biomedical questions. Extracted information as evidence could be applied in improving healthcare and be used by healthcare decision makers, which is the principle of EBM. Extracted information could contain terms that are keywords and keyphrases. Therefore, the specific research questions in this project are: is it possible to explore evidence using term extraction techniques? Can term extraction techniques be used to find clinical answers? What extraction techniques are most appropriate to identify parts of the answer? In order to explore these questions, we implement selected terms extraction to exploit terms which could be hidden knowledge and could identify clinical questions. As the terms are topical words and phrases, the second objective of this project is to use those terms as core concepts for the starting point to compose the abstractive summarisation. This research project makes a step towards achieving abstractive summarisation.

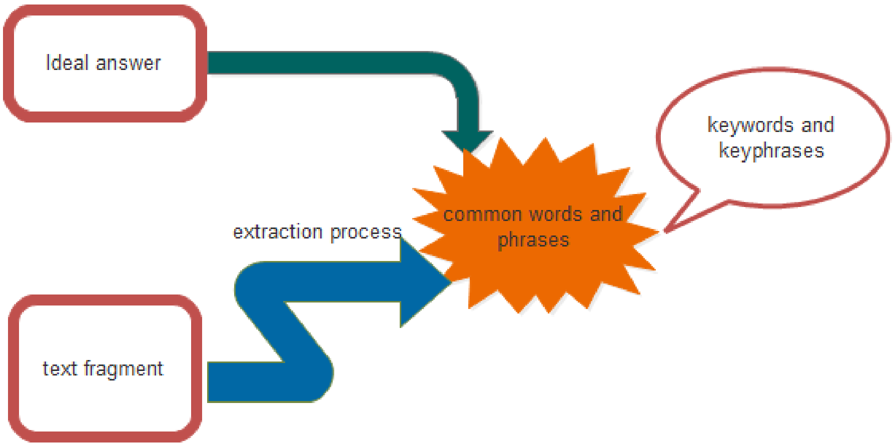


FIGURE 1.6: Searching process of keywords and keyphrases

1.6 Methodology

This research project has two stages. The first stage of this project is to choose extraction techniques to extract terms from relevant answers. The second stage is to determine the most effective technique using a series of outputs in comparison of words in "ideal_answer".

There are a number of different possible extraction approaches which can explore BioASQ data. We review a series of state-of-the-art keyword and keyphrase extraction techniques and use selected extraction approaches to extract keywords and keyphrases. They are TF-IDF, noun phrases filter and C/NC-values. We also propose the neighborhood keywords extraction approach by observing the testing dataset. Thus this research project uses BioASQ dataset, and implements selected approaches individually. These automatic extracting approaches are modified for testing BioASQ data and are domain independent. Without domain knowledge, our project aims to find as many topical terms as possible. This process is described in Figure 1.6. All extraction approaches are repeated several times to test reliability. We investigate how terms number influence term extractions performance and what changes are made to the performance. In addition, these selected approaches are evaluated using "ideal_answers" in BioASQ. The common terms between extracted terms and "ideal_answers" could be answers. Finally, we can determine the most effective extraction approach to answer biomedical questions in the BioASQ dataset.

1.7 Thesis structure

In this chapter, we have introduced the goals and motivation of this research project, which is to determine the most appropriate extraction approach to extract relevant terms as the answers. This thesis is structured as follows. Chapter 2 reviews the previous related work and current keywords and keyphrases extraction techniques. Chapter 3 describes the BioASQ project and preprocessing of the data. Chapter 4 describes the selection of keywords and keyphrases methods, and discusses the process of implementing the methods for BioASQ data, including the problems and how they were resolved. Chapter 5 provides the evaluation framework and evaluates the results. Chapter 6 summarises the research project and its contribution, and presents future research directions.

2
Literature review

2.1 Introduction

Most human knowledge is recorded in text: text is the largest store of human knowledge and it continues to increase. Due to the large and increasing volume of information stored as text written in a natural language, there is a strong need to extract hidden knowledge from text. Extraction information can be used in query-based summarisation. The summarisation technique include extracts and abstracts. Hovy & Lin (1998) reported "To create extracts, one needs procedures to identify the most important passages in the input text. To create abstracts, the core procedure is a process of interpretation". In this research project, we use extracts as the summarisation to answer clinical questions.

The automatic information extraction of medical literature is a timely undertaking task. Specifically, the goal of biomedical text mining research is to identify needed information more efficiently, uncover relationships obscured by the sheer volume of available information, discover knowledge and put it into practical use in the forms of diagnosis, prevention and treatment (Cohen & Hersh, 2005). These existing extraction approaches could help researchers to summarise medical applications and clinical practice. This chapter contains a review of extraction approaches and other related topics: Section 2.2 describes existing extraction systems, Section 2.3 focuses on the related works which are different aspects of biomedical text mining tasks and Section 2.4 reviews a range of keywords and keyphrases extraction techniques which are then selected in a series of experiments in Chapter 4.

2.2 Existing extraction systems

Extraction systems are used to extract factual information and are mainly divided into three types: rule-based, dictionary-based and machine-learning.

Rule-based systems are typically inexpensive in computing, but they need rules created by human experts or linguists. These rules are lexicon rules, pattern rules, semantic rules, formatting rules and hypothesis rules. Rule-based systems classify data by certain string and syntactic characteristics, morphological and orthographic. For example, pattern rules can extract the format which matches the symbols appearing in text, such as the email address xxx@domain and the Cascading Style Sheets form $< ID >$something$< /ID >$. To recognise a variety of strings and formats, rule-based systems need to add more rules to address certain patterns. Lexical rules can specify certain terms such as titles, entities and names of celebrities. However, these rules only perform in certain fields and are not able to develop new rules to solve increasing complexities in text. Furthermore, a list of rules is specific to a certain domain such as medicine and science, and it is difficult to extend to a new domain. Manual setting of rules for a knowledge library is not efficient and time-consuming. In other words, it is difficult to automatically obtain rules and it requires significant manual work.

Dictionary-based systems overcome the general limitations of rule-based systems. Dictionary-based systems mainly make use of the collection of existing terms to identify unknown terms. Such systems need a manual dictionary created to extract information. However, due to the growth of biomedical terminology, a vast number of biomedical terms and entities may have different spellings and forms such as protein names and genome names. This leads to a very low precision and recall (precision and recall are discussed in Chapter 5). For example, the terms prostaglandin-endoperoxide synthase 2, cyclooxygenase-2 and COX2 all point to the same enzyme but have different spellings. Cohen (2005) "it is unclear how these systems will handle genes discovered after system training is complete" and he advised using online resources about genes to construct the dictionary. However, the main issue is that synonyms and abbreviations are continually expanding, and it is impossible to keep up-to-date with developing systems. Normally, the dictionary-based system still needs to be combined with other techniques and human intervention in extracting tasks. There are a range of annotated corpora based on training data to help build the dictionary, especially for finding proteins such as Yapex (Franzén et al., 2002), BioCreative (Hirschman et al., 2005), GENIA (Kim et al., 2003), and GENETAG (Tanabe et al., 2005).

Machine-learning systems use a set of features through training data which is from textual content. Features are examined by machine-learning systems using annotated datasets and machine-learning systems develop models being affected by annotated data to access new data. The advantage is typically that the information of interest is not specified explicitly by the users and, instead, the systems provide a set of documents that have been found to contain the characteristics of interest (the positive training set), and another set that does not (the negative training set) (Cohen & Hersh, 2005), which is used for classification.

However, the machine-learning system is expensive to produce using the annotated corpus and needs a large amount of data which is expensive to obtain or can be out-of-date. In addition, the machine-learning system needs the annotated data to train models and infer new features. But the models could face overfitting problems where the mature system is trained by particular features and cannot be used successfully on new data. In this research project, there are no annotated data or pre-trained systems available.

2.3 Related medical text mining research

EBM has been defined as the process of systematically finding, appraising, and using contemporaneous research findings as the basis for clinical decisions (EBM-Group, 1992). Although EBM aims to search the best available evidence from research publications with benefits to bridge the clinical practice gap, researchers and medical practitioners still have difficulties obtaining specific information for medical practice and medical decision. Besides EBM, other related research on medical text mining focuses on named entity recognition, synonym and abbreviation extraction, and biomedical event extraction. Text classification and relationship extraction in medical text mining are not related to our project.

2.3.1 Named entity recognition

A named entity is a term or phrase that identifies an object from a set of the same items that have similar attributes or the same classification. Such entities categorise people, geographic location, company names, addresses, drug names and protein or gene names. Larkey et al. (2003) conducted a study that showed the importance of proper names in the information retrieval task using named entity recognition. For example, given a query “FIFA” or “World Cup 2014” as the input, the search engines return related named entitis, like ‘Brazil team’ and ‘Lionel Andrés Messi’, which is an interesting association of inputs.

In biomedical text, biomedical named entity recognition can identify cells, DNA, RNA, proteins and genes from the biomedical literature, as well as identify categories. One of the successful named entity recognition systems for biomedical text exploring is AbGene (Tanabe & Wilbur, 2002), for extracting genes and proteins. With Brill POS (Part-of-speech presented in section 2.4.3) tagger (Brill, 1992), it has been trained on 7,000 hand-tagged sentences from biomedical text. The system reaches a precision of 85.7% and at a recall of 66.7%. Precision is the number of correct terms in an extracted set and recall is the number of correct terms divided by the positive terms in text. We discuss precision and recall in detail in Chapter 5. Besides AbGene, there are public named entity recognition resources such as, GENETAG (Tanabe et al., 2005), MEDstract (Pustejovsky et al., 2002), Protein Design Group (PDG) (Franzén et al., 2002) and University of Wisconsin (Blaschke et al., 1999).

2.3.2 Synonym and abbreviation extraction and ambiguity

Synonym and abbreviation extraction are more challenging tasks because most biomedical entities and terminology have multiple names such as gene and protein names. Investigators have used a dictionary of biomedical terms or biomedical synonym lists to explore them. Friedman & Liu (2002) made use of synonym and abbreviation lists by MEDLINE. The evaluation showed a precision of 96.3% and a recall of 88.5%. By manual assigned rules, Yu et al. (2002) obtained a precision of 95% and a recall of 70%. Cohen (2004) applied automatic extraction techniques for MEDLINE abstracts and a numeric analysis metric, and he found a 10% increase in recall with inferring synonyms than without inference.

On the other hand, ambiguity in the context of biomedical articles is very common, especially term ambiguity (Holzinger et al., 2014). There is a tool named MetaMap (http://mmtx.nl m.nih.gov) and it is a disambiguation model, which is a highly configurable program developed by Dr. Alan (Lan) Aronson. MetaMap maps biomedical text to the Unified Medical Language System (UMLS, http://www.nlm.nih.gov/resear-ch/umls/) and matches Metathesaurus concepts referred to in text. UMLS provides a representation of biomedical knowledge consisting of concepts classified by semantic type and both hierarchical and nonhierarchical relationships among the concepts (Aronson, 2001). UMLS is also the largest vocabulary in biomedical domain. MetaMap and UMLS are the main source of explaining terms to avoid ambiguity in biomedical domain using natural language processing and computational-linguistic techniques. The outputs of MetaMap are human-readable and can be generated in the format of XML. Although MetaMap solved ambiguity, there is still room to improve.

2.3.3 Biomedical event extraction

The concept of biomedical event extraction was proposed in 2009. The biomedical event extraction is used to refer to the task of extracting descriptions of actions and relations among one or more entities from the biomedical literature (Kim et al., 2009; Vlachos & Craven, 2012). This approach selected nine events from GENIA (http://www.geniaproject.org/). They are Gene_expres-sion, Transcription, Protein_catabolism, Phosphorylation, Localization, Binding, Regulation, Positive_regulation and Negative_regulation. The event consists of a trigger and at least one argument (protein or events). Protein names which are annotated in advance are tokenised in sentences to trigger the nine event types.

Shared Task on Event Extraction is the first large scale evaluation system, including 13,623 manual annotated terms in 1,210 PubMed citations abstracts. In 2009 and 2011, Kim et al. (2009) and Björne et al. (2009) demonstrated the improvement in performances and the evaluation of F1-metric achieved. F1-metric has gotten 51.95 % and 57.46 % separately. F1-metric is the comprehensive evaluation of precision and recall presented in Chapter 5. From now on, the biomedical event extraction is open towards the more general medical domain and can process the full text rather than abstracts. Biomedical event extraction can be combined with machine-learning and rule-based techniques. As unsupervised learning, Bui

& Sloot (2012) implemented a rule-based approach which set manual rules. They achieved high precision but very low recall. As supervised learning, Björne et al. (2009) treated event extraction as a series flowchart: preprocessing, event trigger word identification, event element identification, and end process. Another event extraction is MineEevent, proposed by Miwa et al. (2012). It improves the whole process, adding more features and including disambiguation to solve co-occurrence.

2.4 Keywords and keyphrases extraction

Keywords and keyphrases form a quick index for a large body of literature. For example, scientific articles and research publications are assigned keywords and keyphrases by the author(s). In online documents, keywords and keyphrases can present the main idea of the content, tag the key points and link to other online resources. Furthermore, keywords and keyphrases provide a high quality description of the given document. They not only provide the basic idea of the documents, but also help readers to search for further details more effectively or decide whether or not to continue further reading. Identifying keywords and keyphrases from documents can help the process which summarise contents explicitly, rapidly and concisely. In principle, they could be used as the core part of abstractive summarisation, forming the relevant information to a query. Dostál & Jezek (2011) reported the keyword extraction and keyphrase extraction:

- Individual keyword extraction — individual words with a special and important meaning, generally in the form of a noun or named entity.

- Keyphrases extraction and derivation — phrases contain two or more keywords and other information in a human-readable form. This phrase can contain verbs and stop words for better readability. Short phrases can be joined by their co-occurrence percentage number.

However, manual assigned keywords and keyphrases are extremely expensive, time-consuming and inefficient to implement. The alternative solution is to automatically extract keywords and keyphrases, and the goal of extraction is to produce topical words or phrases. Automatic extracting techniques are widely used in the internet, databases and documents. The purpose of our research project is to extract keywords and keyphrases from “text” in BioASQ dataset, in order to use them as the answers of clinical questions. Based on these, it is highly desirable to concentrate on the most appropriate extraction techniques for our testing dataset. These existing extraction approaches mainly include simple statistics approach, linguistics approaches, machine learning approaches and other approaches (Giarlo, 2005).

2.4.1 Simple statistics approaches

Early extraction work focused on the frequency of terms in the document. The approach is to directly count the frequency of words and is simple to implement. High frequency terms can be extracted from text as keywords and keyphrases in a single document. The advantage of this method is it does not need training data and linguistic features. But it is impossible to calculate and distinguish important words and phrases in multiple documents. Later on, term frequency research shifted to multiple documents using term frequency — inverse documents frequency (TF-IDF). Lott (2012) described "TF-IDF does this by weighting the term positively for the number of times the term occurs within the specific document, while also weighting the term negatively relative to the number of documents which contain the term". To recognise the keywords using statistical approaches, we consider and select TF-IDF as our first experiment in our project as discussed in Chapter 4.

N-gram is the language model for extracting words sequence. The most popular N-gram application categories include: Unigram, Bigram and Trigram. A Unigram only takes one term at a time, a Bigram takes two terms at a time, and so on. N-gram is the probability of the whole sentence by multiplying the probability of every word which occurred in the sentence. The practical model of N-gram is a probabilistic language model and is a form of the (N-1) Markov model — the Nth term is only dependent on the (N-1)th terms. The disadvantage is it requires extensive computer processing power using N-gram with Markov chain. Furthermore, N-gram can be used in recognising phrases. Qiu et al. (2012) said "if a word sequence is better modeled by an N-gram language model than by a Unigram language model, then it is more likely to be a phrase".

Word co-occurrence is another statistical extraction method proposed by Matsuo & Ishizuka (2004). Two terms which appear in the same sentence are considered to co-occur. Words are important terms if they co-occur in the document more frequently than if they are randomly distributed in the document. However, the clear problem is a term frequency is too low or if a word only occurs in the document, which cannot support statistical significance due to extremely sparse data (Lott, 2012). To overcome this problem, the authors used the χ^2 values to determine the biases between expected and observed frequencies.

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g}$$

The term g∈G, the expected probability p_g and frequent terms G as n_w, the total number of co-occurrence of term is w. The term $n_w p_g$ represents the expected frequency of co-occurrence and $(freq(w,g) - n_w p_g)$ represents the difference between expected and observed frequencies (Matsuo & Ishizuka, 2004). However, the authors have not illustrated clearly the expected frequency and how to define the threshold. They reported the co-occurrence was able to closely match TF-IDF, but did not rely on the use of a document corpus (Matsuo & Ishizuka, 2004; Lott, 2012).

2.4.2 Machine learning approaches

Machine learning requires data, especially high quality data. Data can be created by the previous history of the optimisation process or by feedback from the decision maker (Battiti & Brunato, 2014). The data is divided into three sets: a training set (to label the examples and shape the model), a validation set (to evaluate the model) and a test set (to get results). Machine learning approaches use the training dataset to be guided. In supervised learning, the system is trained by labelled examples. Through the examples as the vector of input parameters called features, the system will develop the ability to recognise correct examples — words and phrases being extracted as keywords and keyphrases. Hulth (2003) used supervised learning paradigms to determine the best keywords using within-document frequency, collection frequency, relative position of the first occurrence and sequence of part of speech tagging (discussed in Section 2.4.3), which are obtained from both the training data and test data.

Keyphrase Extraction Algorithm (KEA) system automatically extracts keyphrases from text, using lexical methods and machine-learning algorithms to predict which candidates are good keyphrases (Witten et al., 1999). The algorithm is obtained in the feature model from the training documents where the keyphrases annotated by the author are identified. KEA uses TF-IDF and first occurrence as the features, where both features are numbers to translate machine learning scheme. KEA considers the candidate phrases with the two features against the keyphrases by the author, then implements the new document. The system uses the model to choose keyphrases from a new document. The results show that KEA can extract keyphrases which are partly assigned by the author, but some KEA keyphrases are poor compared to keyphrases by author(s).

In our case, this research project does not have the training data for experiments. But, machine-learning is still an important direction. Besides classification techniques in machine-learning systems discussed in Section 2.2, there are other machine learning techniques, including support vector machines (SVM), conditional random fields (CRF), hidden Markov model (HMM), Decision tree model, naïve Bayesian, and Multi-layer perception (MLP).

2.4.3 Part of Speech approach

Many text mining projects make use of linguistic resources to extract, compare and evaluate the outputs. Part of speech (PoS) tag pattern is one of the linguistic approaches for assisting extraction. It is also called grammatical tagging or word-category disambiguation. PoS tag pattern normally assists a term selection task, being a linguistic category of syntactic properties. Table 2.1 shows a sample of PoS tagging. In each pair of brackets, the first element is the word and the second element is the Part of Speech. Most of the second elements in lexical categories are mostly nouns, pronouns, adjectives, verbs, adverbs, conjunctions, prepositions, and interjections. The PoS tag pattern is used in disambiguation. For example, many English words can be more than one part of speech. The task faced by disambiguation is to extract a word “run” which has both “verb” and “noun” properties in a sentence. For referring a particular type, PoS Tag pattern can help explicitly determine which properties could be selected with lexical analysis. Also PoS tag pattern can make use of statistical methods to display the number of similar behaviour terms. Hulth (2003) reported that adding linguistic knowledge to the representation (such as syntactic features), rather than relying only on statistics (such as term frequency and N-grams), could get better results. He gave a good overview of linguistic and statistical methods working together. This idea uses the statistical features which are proposed by Frank et al. (1999) and Hulth (2003). They selected three selection approaches to test — N-gram, NP-chunks (presented in the following part) and PoS tag pattern. As judged by the validation set, not every approach with features is better than the manually assigned keywords. However, by adding PoS tag pattern, the test outperformed itself without PoS tag pattern and the test has achieved NP-chunks results better than N-gram.

Table 2.1: Sample of Part of Speech tagging

```
[('Disease', 'NN'), ('patterns', 'NNS'), ('in', 'IN'), ('RA', 'NNP'),
('vary', 'NN'), ('between', 'IN'), ('the', 'DT'), ('sexes;', 'NNP'),
('the', 'DT'), ('condition', 'NN'), ('is', 'VBZ'), ('more', 'RBR'),
('commonly', 'RB'), ('seen', 'VBN'), ('in', 'IN'), ('women,', 'NNP'),
('who', 'WP'), ('exhibit', 'NN'), ('a', 'DT'), ('more', 'RBR'), ('ag-
gressive', 'JJ'), ('disease', 'NN'), ('and', 'CC'), ('a', 'DT'), ('poorer',
'NN'), ('long-term', 'JJ'), ('outcome.', 'NNP')]
```

There are two famous lexical resources: the Electronic Dictionary Research (EDR) electronic dictionary and WordNet. Although they are not extraction techniques, they are useful standards in terms extractions. EDR is a dictionary for computer processing lexical information of natural language, such as word processing and machine translation. It captures lexical information, such as lemmas (discussed in Chapter 3), part of speech, and word senses to support the levels of morphology, syntax, semantic and pragmatics (Suematsu, 1994). WordNet from Princeton University is also freely available online and provides super-subordinate relation, the part-whole relation, Verb synsets and Cross-PoS relations. WordNet is a network of words by semantic relationship and linking their concepts.

Besides PoS tag, Kaur & Gupta (2010) reported that linguistic approaches included lexical analysis, syntactic analysis and discourse analysis. In this research project, we select C/NC-values as one of our experiments in Chapter 4, which returns potential keyphrases using both statistical and linguistic information. Specifically, C/NC-values is one of the extraction approaches, which combines measure of frequency of occurrence and sensitivity to a particular type of multi-word terms. The higher the C/NC-values scores, the greater chance phrases are keyphrases.

2.4.4 Noun Phrase-chunks approach

Chunks split a sentence into pieces with semantic parts rather than into individual words and symbols. The process is to group the syntactically related words into the same phrase and these phrases are divided into non-overlapping phrases within a sentence. It is expected chunks produce more real phrases than N-grams and a chunk recognises the N-grams process having a certain linguistic property (Hulth, 2003). Qiu et al. (2012) pointed at that "we use a noun phrase chunker to identify noun phrases and filter out those N-grams that are not noun phrases". The chunk types are varieties such as verb phrase (VP), adjective phrases (ADVP), adverb phrase (ADVP) and preposition phrase (PP). Figure 2.1 shows examples of Noun Phrases-chunks, The smaller boxes show the word-level split (tokenisation) and part-of-speech tagging, while the large boxes show higher-level chunking where these larger boxes are called a chunk (Bird et al., 2009).

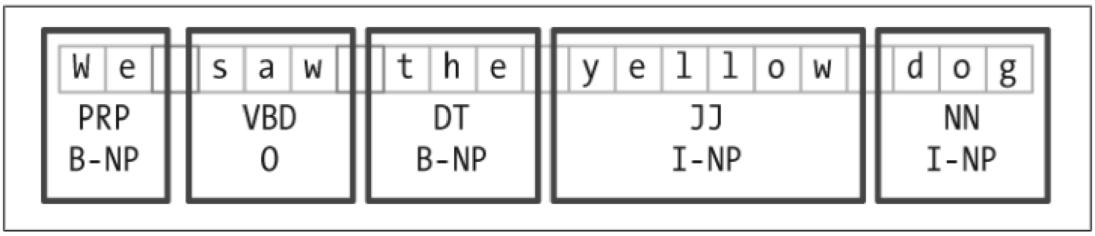


Figure 2.1: NP-chunk
Source : http://www.nltk.org/book/ch07.html

Abney (1991) proposed the advantage of chunking parsers and how to chunk in context by strategies, which is the foundation level of analysis. As text chunking consists of syntactic constituents without any further embedded constituents, the problem for the human parser is whether to adopt the chunk-by-chunk strategy. Ramshaw & Marcus (1999) combined the chunking techniques using machine learning techniques. They started the supervised training corpus, a set of rules and a baseline heuristic, then the pattern matched the particular features. Later on, Veenstra & Buchholz (1998) used the dataset described in Ramshaw & Marcus (1999) with memory-based learning to quick chunk. Now most efforts by researchers rely on ambiguities.

NP chunks as indicative keyphrases can summarise a document more concisely and aid the topic search. Wan & Xiao (2008b) reported "appropriate keyphrases can serve as a highly condensed summary for a document, and they can be used as a label for the document to supplement or replace the title or summary". In our research project, we select three NP patterns to extract keyphrases in Chapter 4. In addition, NP chunks could be used in a series of natural language processing and information retrieval tasks, such as citation summarisation (Qazvinian et al., 2010), text summarisation (Hulth & Megyesi, 2006), opinion mining (Berend, 2011), document indexing (Gutwin et al., 1999), document classification (Krulwich & Burkey, 1996), and document cluster (Hammouda et al., 2005).

2.4.5 Informative features approach

Words have various forms in documents which are hints to emphasise additional, but substantial information about the importance of the words. Alguliev & Aliguliyev (2005) summarised informative features for keyword extraction such as:

- Words emphasised by application of bold, italic or underlined fonts
- Words typed or written in upper case
- The size of the font applied

2.4.6 Position features approach

Kaur & Gupta (2010) demonstrated that words in different positions carry different entropy (information theory, entropy is a measure of the uncertainty in a random variable. In this context, the term usually refers to the Shannon entropy, which quantifies the expected value of the information contained in a message). They revealed candidate words that appeared in the titles, abstracts, introductions and summary paragraphs were more valuable than those words that appeared in the body and references. In such valuable positions, they carry more critical information than in ordinary paragraphs. For example, if the same word appears in the title and conclusion, that word carries more weight. They are more likely to be keywords and keyphrases in these positions. The candidate keywords and keyphrases in different positions can be applied by different weights when scanning. For example, we can assign the title word 5 points, the conclusion word 3 points and the abstract word 2 points. We can use the weights to distinguish the importance of keywords and keyphrases in the given article. The advantage of position features are that users can easily find keywords and keyphrases.

2.4.7 Query-based extraction for summarisation

Query-based extractive summarisation targets description of questions and returns the summary with relevance to the questions. Query-based extractive summarisation adds structure to documents and grammatical sentences. Thus, in some sense, query-based extraction is task-specific and user oriented. The outputs of query-based extraction are more or less directly readable by users. Chowdary et al. (2010) proposed summarisation using graphs. In their paper, the contextual relationships are exploited and sub-graphs are constructed, which consists of highly relevant sentences to the query. They used the scoring model to rank and select the highest ranked sub-graph as the summary.

Jin et al. (2009) described the five components of query-based extractive summarisation: Preprocessing, Question analysis and Query generation, Sentence Scoring, Opinion polarity detection, Redundancy removal and summary generation. In another query specific extractive summarisation, Ma et al. (2008) used the five components process and proposed the query-related features and topic-related features to identify important words in the relevant document. They modified the Maximal Marginal Relevance (MMR) to adjust scores of candidate sentences including important words and choosing the highest score sentence as the summary. They presented "the relevant degree of R(w1, w2) is calculated by taking a window of length K words and moving across the text at one word increments. All words in the window are said to co-occur with the first word with strengths inversely proportional to the distance between them where n(w1, k, w2) is the number of w1 and w2 co-occurring in the window, and k denotes the real distance between w1 and w2 when they are co-occurred." The principles of the keyword extraction formulas are shown below.

The relevant degree R(w1, w2) formula and qwt is the query sentence:

$$R(w1, w2) = {}^{K}\sum_{k=0} w(k)^{*} n(w1, k, w2)$$

and the query-related feature formula:

$$F1(w_i) = {}^{qwt-1}\sum_{j=0} R(w_i, w_j)$$

Besides all of the extraction approaches illustrated above, other useful and novel approaches include: graph-based keyword extraction (Litvak & Last, 2008), keyword extraction by conditional Random Field (CRF) (Zhang, 2008), keyphrases extraction by neural networks (Wang et al., 2006; Sarkar et al., 2010), TextRank (Mihalcea & Tarau, 2004), ExpandRank (Wan & Xiao, 2008b), and SingleRank (Wan & Xiao, 2008a). But some approaches are not suitable due to our dataset type constraint. We cannot cover all of them in this project due to time constraints and page limits of this thesis.

2.5 Summary

Current literature often focuses on how to extract terms concisely instead of reorganising text. So far, it is still not clear how to generate an abstractive summarisation, but we can develop some general guidelines and effective approaches to move from extracts to abstracts. Our research project uses keywords and keyphrases extraction to construct the abstract notation because keywords and keyphrases extraction is the most popular strategy for query-based summarisation, and it computes the importance of candidate keywords and keyphrases. Eventually, the techniques of abstractive summarisation include natural language processing, semantic analysis, information retrieval, domain knowledge and statistical approaches, which could be used in a variety of domains and genres.

In this chapter, we have reviewed previous research on medical text, exploring a range of extraction techniques. These techniques can be used in the medical domain. We focused on the state-of-the-art techniques based on statistical approaches, linguistics approaches, machine learning approaches and other approaches in the literature review. Chapter 3 presents the background of BioASQ and its dataset, and preprocessing of data. Based on the review in this chapter, we choose selected approaches to test our dataset, discussed in more detail in Chapter 4.

3

Data and Preparation

3.1 Introduction

In Chapter 1, we briefly presented the BioASQ project and aims of this research project. In this chapter we introduce details of BioASQ project and data preprocessing. In Section 3.2, we outline the BioASQ project and its activities. In Section 3.3 we discuss the BioASQ dataset where we develop our methods to extract information. In Section 3.4 we show the two main software tools used: Natural Language Toolkit (NLTK) which is a natural language processing software package of Python, and scikit-learn which is a machine learning software package of Python. In Section 3.5 we illustrate the preprocessing to avoid an unacceptable level noise so that we can successfully access the dataset.

3.2 BioASQ Project

As stated in Chapter 1, BioASQ is an European project which provides data, software and evaluation infrastructure to explore biomedical text. The project ensures that biomedical experts in future can rely on software tools to identify, process and present the fragments of the huge quantity of biomedical resources that address biomedical questions (Paliouras & Krithara, 2015). The challenge of BioASQ has two tasks: BioASQ Task on Large-Scale Online Biomedical Semantic Indexing and BioASQ Task on Biomedical Semantic question-answering. Our testing dataset is collected from the latter task, which is task B. The dataset used is a question-answer set and provides related information. Our project extracts terms from text fragments that are the benchmark dataset containing developments. Text fragments are mostly from abstracts in MEDLINE and they are from relevant articles, snippets of the articles, relevant concepts and Resource Description Framework (RDF) triples from linked life data. These developments are potential answers which are possible to identify in this type of clinical question.

The BioASQ challenge includes research interests such as classification, document retrieval, fact checking, information extraction, information retrieval, machine learning, named entity disambiguation, name entity recognition, natural language generation, passage retrieval, Question Answer (QA) from structured information, QA from Text, reasoning, RDF Triple, relation extraction, semantic indexing, text summary, and textual entailment (BioASQ, 2014a). The participants take the two tasks with interests to index journal abstracts and retrieve related information for test questions which are described in English. BioASQ also offers several sub-tasks to participate in such as, retrieving PubMed documents that contain an answer, retrieving snippets from those documents that contain answers, retrieving relevant concepts, and extracting the answer from all retrieved material (Lingeman & Dietz, 2014).

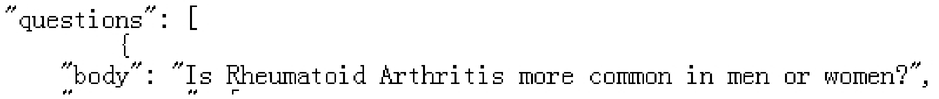
3.3 BioASQ dataset

The version of BioASQ dataset used in this study was released in 2013. The dataset is a question-answering document collection, providing biomedical questions and relevant text fragments with their properties and sources. Text fragments are collected responses that could be values to the question. For example, if the question is "*What is the function of the mammalian gene Irg1?*", there are several possible related fragments such as "*we identified three interferon-stimulated genes (ISGs; I27, Irg1 and Rsad2 (also known as Viperin) that mediated the antiviral effects against different neurotropic viruses*", and "*The proinammatory cytokine-induced IRG1 protein associates with mitochondria*".

In addition, our testing dataset contains 310 question units and every unit includes a "body" which is a question, "snippets", "ideal_answer", "exact_answer" and properties. The properties include ID, a description of concept and related information. As an example, the real question in one unit is presented in Figure 3.1. The full content of a unit is given in Appendix A.2. Figure 3.1 to Figure 3.4 are real examples from BioASQ dataset.

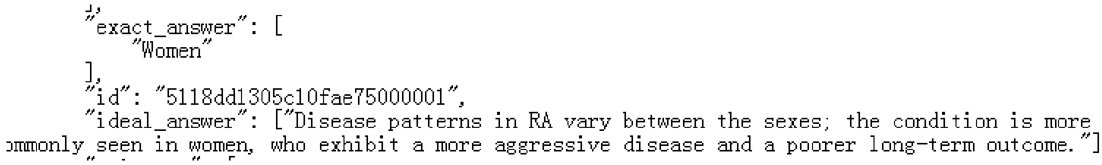
Figure 3.2 below presents examples of "exact_answer" and "ideal_answer". The "exact_answer" is the names of particular diseases, symptom and genes. The "ideal_answer" is a paragraph-sized summary with the information most important to a question. Table 3.1 shows the questions with their "exact_answer" and "ideal_answer". The two types of answers are gold answers representing the best answers and they are provided by biomedical experts.

The body of each "snippet" has several text fragments with their properties arranged by groups. These "'text" are relevant answers and are potential explanations to the biomedical question from publications. Figure 3.3 presents the part of "snippets". In our project, since



```
"questions": [
        {
    "body": "Is Rheumatoid Arthritis more common in men or women?",
```

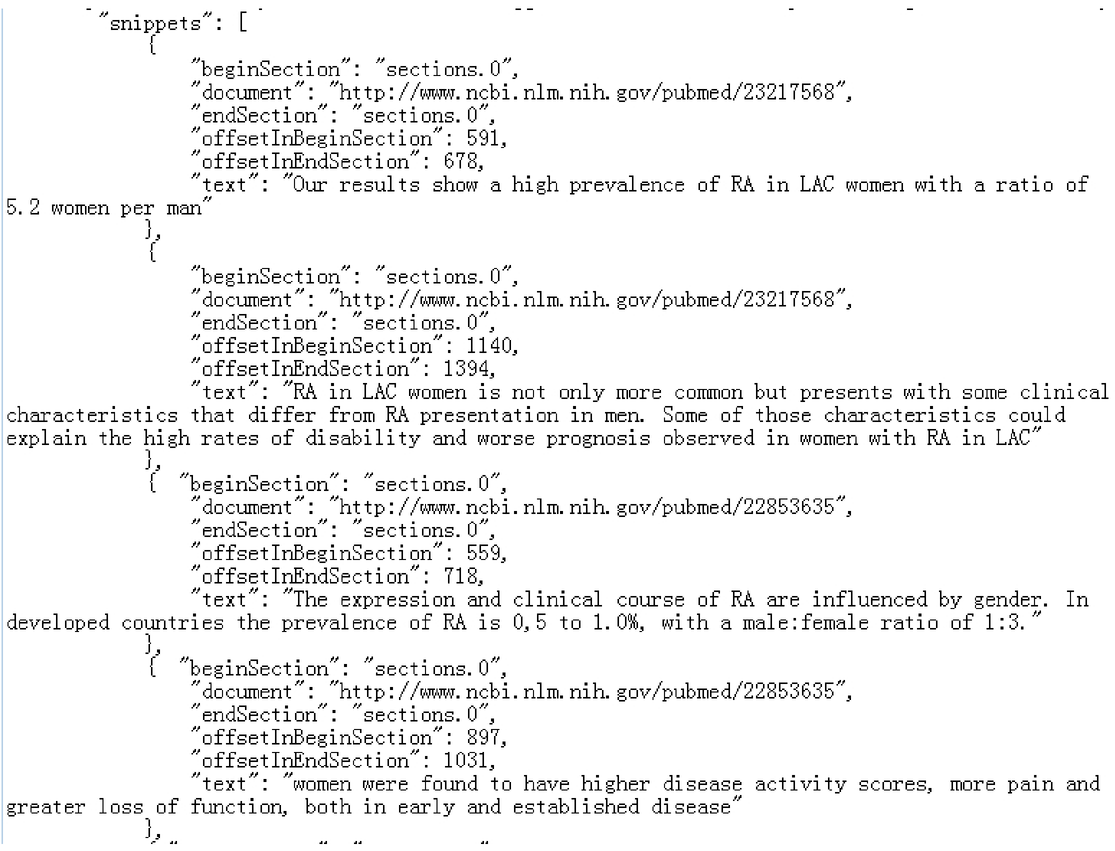
Figure 3.1: Sample of a BioASQ question



```
},
"exact_answer": [
    "Women"
],
"id": "5118dd1305c10fae75000001",
"ideal_answer": ["Disease patterns in RA vary between the sexes; the condition is more
ommonly seen in women, who exhibit a more aggressive disease and a poorer long-term outcome."],
```

Figure 3.2: “Exact_answer” and “ideal_answer” to question in Figure 3.1

a group of “text” are related responses for one question, we are interested in the “text” in the “snippets” so that we can explore keywords and keyphrases. Keywords and keyphrases are automatically evaluated against “ideal_answer” that is the evaluation standard in Chapter 5.



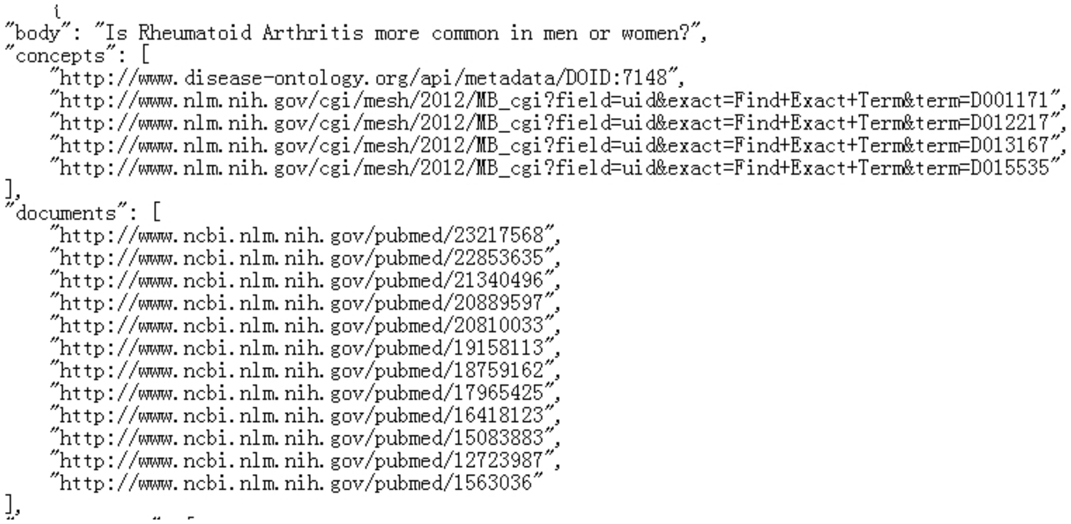
```
    "snippets": [
        {
            "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
            "endSection": "sections.0",
            "offsetInBeginSection": 591,
            "offsetInEndSection": 678,
            "text": "Our results show a high prevalence of RA in LAC women with a ratio of
5.2 women per man"
        },
        {
            "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
            "endSection": "sections.0",
            "offsetInBeginSection": 1140,
            "offsetInEndSection": 1394,
            "text": "RA in LAC women is not only more common but presents with some clinical
characteristics that differ from RA presentation in men. Some of those characteristics could
explain the high rates of disability and worse prognosis observed in women with RA in LAC"
        },
        {  "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
            "endSection": "sections.0",
            "offsetInBeginSection": 559,
            "offsetInEndSection": 718,
            "text": "The expression and clinical course of RA are influenced by gender. In
developed countries the prevalence of RA is 0,5 to 1.0%, with a male:female ratio of 1:3."
        },
        {  "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
            "endSection": "sections.0",
            "offsetInBeginSection": 897,
            "offsetInEndSection": 1031,
            "text": "women were found to have higher disease activity scores, more pain and
greater loss of function, both in early and established disease"
        },
```

Figure 3.3: Text in the “snippets”

TABLE 3.1: Types of questions with gold answers (from the testing dataset)

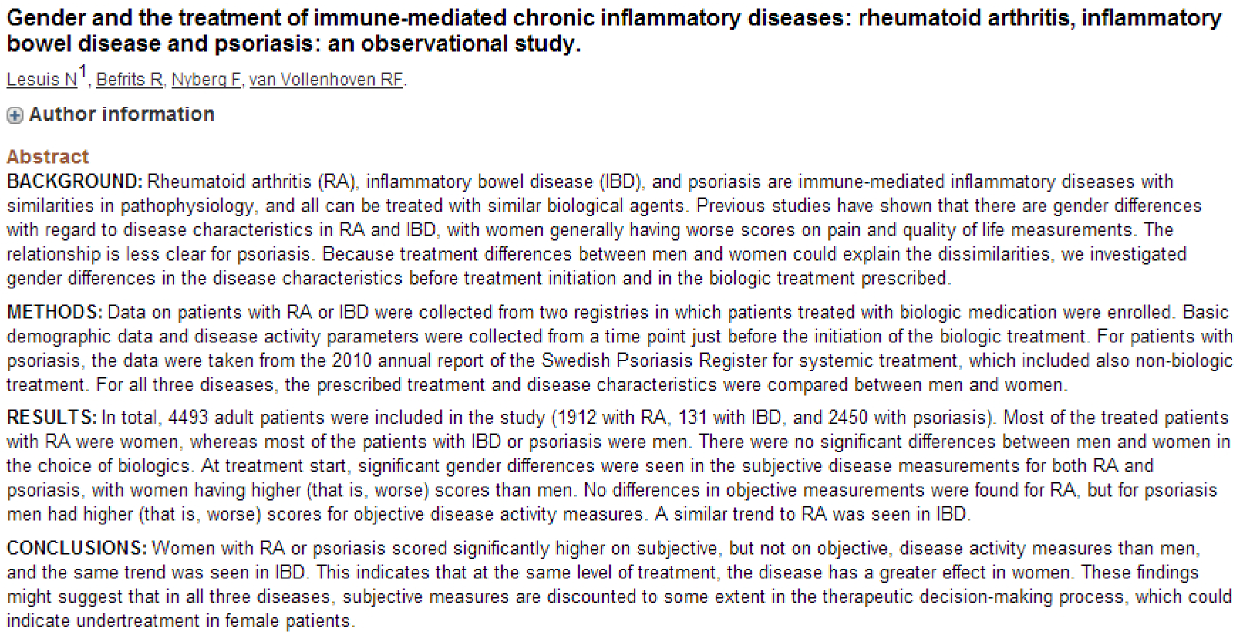
Question type	**Required answer**	**Example question**	**Exact answer**	**Ideal answer**
Yes/No	Exact + Ideal	Are there any DNMT3 proteins present in plants?	Yes	Yes. The plant DOMAINS REARRANGED METHYLTRANSFERASE2 (DRM2) is a homolog of the mammalian de novo methyltransferase DNMT3. DRM2 contains a novel arrangement of the motifs required for DNA methyltransferase catalytic activity.
Factoid	Exact + Ideal	Which is the neurodevelopmental disorder associated to mutations in the X- linked gene mecp2?	Rett syndrome	The neurodevelopmental disorder named Rett syndrome, originally termed as cerebroatrophic hyperammonemia. Although most exclusively affects females, has also been found in male patients.
List	Exact + Ideal	Which species may be used for the biotechnological production of itaconic acid?	[[“Aspergillus terreus”], [“Aspergillus niger”], [“Ustilago maydis”]]	In 1955, the production of itaconic acid was firstly described for Ustilago maydis. Some Aspergillus species, like A. itaconicus and A. terreus, show the ability to synthesise this organic acid and A. terreus can secrete significant amounts to the media. Itaconic acid is mainly supplied by biotechnological processes with the fungus Aspergillus terreus. Cloning of the cadA gene into the citric acid producing fungus A. niger showed that it is possible to produce itaconic acid also in a different host organism.
Summary	Ideal	What is the main mechanism by which human papillomavirus proteins E6 and E7 contribute to cell transformation?	–	Although they may have other targets, human papillomavirus proteins E6 and E7 interact with and block the function of p53 and pRb, respectively, therefore deregulating cell cycle and leading to cellular transformation.

Figure 3.4 presents several links and each link in "documents" is the source of the individual part in the "snippets". The "concepts" is related information such as disease ontology. Figure 3.5 shows the content of a link. It contains background, methods, results and conclusion. But most contents are only abstracts in such web pages. Besides these parts, the BioASQ dataset has other descriptions which are not considered in our project such as types, concepts and triples (see Appendix A.2).



```
    {
"body": "Is Rheumatoid Arthritis more common in men or women?",
"concepts": [
    "http://www.disease-ontology.org/api/metadata/DOID:7148",
    "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D001171",
    "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D012217",
    "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D013167",
    "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&term=D015535"
],
"documents": [
    "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
    "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
    "http://www.ncbi.nlm.nih.gov/pubmed/21340496",
    "http://www.ncbi.nlm.nih.gov/pubmed/20889597",
    "http://www.ncbi.nlm.nih.gov/pubmed/20810033",
    "http://www.ncbi.nlm.nih.gov/pubmed/19158113",
    "http://www.ncbi.nlm.nih.gov/pubmed/18759162",
    "http://www.ncbi.nlm.nih.gov/pubmed/17965425",
    "http://www.ncbi.nlm.nih.gov/pubmed/16418123",
    "http://www.ncbi.nlm.nih.gov/pubmed/15083883",
    "http://www.ncbi.nlm.nih.gov/pubmed/12723987",
    "http://www.ncbi.nlm.nih.gov/pubmed/1563036"
],
```

Figure 3.4: Source of fragments in BioASQ dataset



Gender and the treatment of immune-mediated chronic inflammatory diseases: rheumatoid arthritis, inflammatory bowel disease and psoriasis: an observational study.

Lesuis N[1], Befrits R, Nyberg F, van Vollenhoven RF.

⊕ **Author information**

Abstract

BACKGROUND: Rheumatoid arthritis (RA), inflammatory bowel disease (IBD), and psoriasis are immune-mediated inflammatory diseases with similarities in pathophysiology, and all can be treated with similar biological agents. Previous studies have shown that there are gender differences with regard to disease characteristics in RA and IBD, with women generally having worse scores on pain and quality of life measurements. The relationship is less clear for psoriasis. Because treatment differences between men and women could explain the dissimilarities, we investigated gender differences in the disease characteristics before treatment initiation and in the biologic treatment prescribed.

METHODS: Data on patients with RA or IBD were collected from two registries in which patients treated with biologic medication were enrolled. Basic demographic data and disease activity parameters were collected from a time point just before the initiation of the biologic treatment. For patients with psoriasis, the data were taken from the 2010 annual report of the Swedish Psoriasis Register for systemic treatment, which included also non-biologic treatment. For all three diseases, the prescribed treatment and disease characteristics were compared between men and women.

RESULTS: In total, 4493 adult patients were included in the study (1912 with RA, 131 with IBD, and 2450 with psoriasis). Most of the treated patients with RA were women, whereas most of the patients with IBD or psoriasis were men. There were no significant differences between men and women in the choice of biologics. At treatment start, significant gender differences were seen in the subjective disease measurements for both RA and psoriasis, with women having higher (that is, worse) scores than men. No differences in objective measurements were found for RA, but for psoriasis men had higher (that is, worse) scores for objective disease activity measures. A similar trend to RA was seen in IBD.

CONCLUSIONS: Women with RA or psoriasis scored significantly higher on subjective, but not on objective, disease activity measures than men, and the same trend was seen in IBD. This indicates that at the same level of treatment, the disease has a greater effect in women. These findings might suggest that in all three diseases, subjective measures are discounted to some extent in the therapeutic decision-making process, which could indicate undertreatment in female patients.

Figure 3.5: One of link descriptions in Figure 3.4
Source: http://www.ncbi.nlm.nih.gov/pubmed/22853635

3.4 Software tools for extractions

Two programming software packages are used for extraction: Natural Language Toolkit(NLTK) (Bird et al., 2009) and scikit-learn (Pedregosa et al., 2011). The NLTK is the software package of Python programming language. It is a high performance, developing, free and open source project for natural language processing. The function of NLTK is a research tool, which can tokenise, PoS tag, identify named entities, and conduct chunking and parsing. It is across platforms and is available for Windows, Mac OS X and Linux. Scikit-learn is an open source library for machine learning by Python programming language as well. It features dimensionality reduction, model selection, preprocessing, regression, clustering and classification. Scikit-learn provides a range of supervised and unsupervised machine learning algorithms and modules such as support vector machine, logistic regression, naïve Bayes, random forests, gradient boosting and k-means. NLTK is able to preprocess text and scikit-learn can help extract keywords using TF-IDF. These software packages help us develop programs to meet the research aims.

3.5 Preprocessing of data

In data mining and text mining tasks, the priority work is to clean and transfer data to an acceptable level because of data containing strings, grammatical separators (“,”,“/”,“;”) or illegal inputs. Special characters and symbols are normally removed as well. The exclusion of case sensitive text is also implemented as they commonly appear in the English language. In our dataset, we keep the numeric characters as they could be important factors in the medical domain.

The required representation of BioASQ dataset cannot be directly accessed by the extraction program we developed. To effectively process the dataset, it is necessary to preprocess the raw data. Preprocessing of data is described below. The four steps are format conversion, tokenisation, stopword removal, and lemmatization and stemming

Format conversion: Text comes in a range of formats such as paper, journals, notes and signal posts and it is only readable by humans. In the digital era, text is normally stored in electrical documents with various formats. However, these formats are different and need to be accessed by different methods, like Microsoft Word and Adobe pdf reader. The users have to install the proper software to access text. For example, complex web pages not only contain text, but also layout framework, such as HTML, Cascading Style Sheets (CSS), or JavaScript. The HTML provides text, images and tables. CSS controls the layout and colour framework, and JavaScript sets the interaction. Due to the HTML structures, the users have to access these elements using different tools. Therefore, all the variables are unified and accessible by design code. In particular, most of the medical lexicon is in Greek or Latin, which are different from the English character encodings, and needs different language code sets such as ASCII, UTF-8 or other character sets. We use the UTF-8 for the encoded characters. BioASQ dataset does not provide pictures or layouts, and the plain text can be directly extracted after filtering noise.

Tokenisation: Text as a representational component must be readable by machines and words in text must be a sequence of characters. This process requires splitting text into sentences and sentences into small pieces to form single words or word-like phrases, which is called tokenisation. Tokenisation can remove white spaces and punctuation marks. By different splitting rules, tokenisation would split a character into different meaning. For example, "isn't" is split into "is" and "n't" or "is", "n" and "t". In our research, we use the default tokenisation of NLTK. The default token is trained on English texts and is for English text processing. One of the practical token tools is Stanford Tokenizer (http://nlp.stanford.edu/softw-are/tokenizer.shtml), which is created by Stanford Natural Language Processing Group.

Stopword removal: Words that are considered non-meaningful for the purpose of text analysis are named stop words. Stop words are syntactic structures in the English language and these words are uninteresting words. They are a set of high frequency words, and they are unnecessary content where these words can be typically filtered out before processing of text. Stop words usually have little lexical content, and their presence in a text fails to distinguish it from other texts (Bird et al., 2009). But there is no common definition of what they are. High-frequency words can include "the", "is", "at", "which", "on", "and", and "to". By properties, these include adverbs, prepositions, conjunctions and punctuations. Previous study has shown that stop words can be either removed by using a predefined list or applying a threshold to the frequency of words in the corpus and removing high-frequency words (Holzinger et al., 2014). The outcome of this step may be smaller than the sequence of tokenisation.

Lemmatisation and stemming: The process of stemming and lemmatising controls the influence of grammatical form in the context. Holzinger et al. (2014) reported that stemming was the process of heuristically removing suffices from words to reduce the word to a commmon form (the word stem), while lemmatisation referred to more sophisticated methods using vocabularies and morphological analysis. The common stemming tool in natural language processing is Porter stemmer (Porter, 1980) and the lemma of biomedical texts is BioLemmatizer (Liu et al., 2012). In English, for example, *measure*, *measured*, *measuring* and *measurement* are forms of the same lexeme, which is called the lemma. This is a very common phenomenon in the English language. We lemmatise all the remaining text after tokenisation. The outcome is still a sequence of terms.

In order to test all extraction techniques we selected, it is necessary to exclude the errors and noisy data so that we can successfully implement various experiments. In our research project, we aim to extract terms by different extraction techniques. However, the most important work is to preprocess the data, otherwise, the program we design is unable to process it. On the other hand, the evaluation is also affected by the source of testing data.

Preprocessing includes tokenising the sentences, removing symbols, transfering upper case to lower case characters and filtering out stop words. Everything else in the sentence remains unchanged. Examples of preprocessing for the first question unit are shown in Table 3.2, 3.3, 3.4 and 3.5. Table 3.2 shows the raw text before the tokenisation. Table 3.3 shows the sentence is divided into the single unit after tokenisation. Table 3.4 shows the text after removal of punctuation and stop words, and transfering to lower case. We use WordNetLemmatizer in NLTK, and the results of lemmatisation removing duplicated words are presented in Table 3.5. In this research project, we use lemmatizstion instead of stemming, because stemming usually removes “e”, “l” and “y” at the end of a word, which breaks the word. Numerical information may contain new knowledge, the preprocessing keeps non-alpha-characters.

Table 3.2: Text before the tokenisation

Disease patterns in RA vary between the sexes; the condition is more commonly seen in women, who exhibit a more aggressive disease and a poorer long-term outcome.

Table 3.3: Text after the tokenisation

[u’Disease’, u’patterns’, u’in’, u’RA’, u’vary’, u’between’, u’the’, u’sexes’, u’;’, u’the’, u’condition’, u’is’, u’more’, u’commonly’, u’seen’, u’in’, u’women’, u’,’, u’who’, u’exhibit’, u’a’, u’more’, u’aggressive’, u’disease’, u’and’, u’a’, u’poorer’, u’long’, u’-’, u’term’, u’outcome’, u’.’]

Table 3.4: Text after removal of punctuation and stopwords, and setting lower case

[u’disease’, u’patterns’, u’ra’, u’vary’, u’sexes’, u’condition’, u’commonly’, u’seen’, u’women’, u’exhibit’, u’aggressive’, u’disease’, u’poorer’, u’long’, u’term’, u’outcome’,]

Table 3.5: Text after Lemmatise the results of Table 3.4

[u’exhibit’, u’term’, u’woman’, u’pattern’, u’outcome’, u’vary’, u’disease’, u’commonly’, u’sex’, u’long’ ,u’ra’, u’seen’, u’aggressive’, u’poorer’, u’condition’]

3.6 Summary

In this chapter, we presented the BioASQ project and its activities. Then we described how we prepared the dataset so that the data could be accessed and useful for our research aims. After the preparation work, the next chapter discusses and implements selected extraction approaches for the BioASQ dataset.

4

Application of extraction techniques to the BioASQ dataset

4.1 Introduction

In Chapter 2, we already reviewed current existing extraction systems and extraction techniques in natural language processing and computational-linguistic. In particular, we focused on term extractions in this project. In Chapter 3, we mainly presented the BioASQ project and its dataset. At the end of Chapter 3, we showed the preprocessing process for BioASQ dataset. To discover potential techniques that can answer questions in BioASQ dataset, this chapter is to implement extraction techniques we selected. The task of term extraction is to identify a series of keywords or keyphrases. To some extent, these terms could be the key information of the summarisation. In other words, they can be answers of biomedical questions in BioASQ dataset.

We have selected five different extraction techniques and we illustrate how these techniques are implemented. In Section 4.2, we test the TF-IDF approach, which is one of the statistical methods. In Section 4.3, we implement the k nearest neighbour words by assigned terms. In Section 4.4, we explore the noun phrases filter which uses linguistic features. In Section 4.5, we test the C/NC-values technique that combines the statistical method and linguistics method to find the terms.

4.2 Extraction approach 1 – TF-IDF

TF-IDF (term frequency–inverse document frequency) is one of the best-known and most commonly used term extraction algorithms (Robertson, 2004). TF stands for term frequency, which is the number of times a specific term appeares in a given document. It can describe occurrences of a term within a single document or multiple documents. Statistically, the higher the frequency, the more it is presumed to be a potentially important term. It is calculated according to the formula:

$$tf(t,d) = \frac{f_{(t)}}{n}$$

- *tf(t,d)* is the amount of a term frequency
- n is the whole number of terms in the given document
- d is the given document
- *f(t)* is term frequency

However, the less meaningful terms which are more frequent can be *a*, *an*, *that*, *of*, *and*, *this*, *in*, *with*, *not*, *only*, and *more*. They are uninteresting words which are very common in documents. In the data preprocessing stage, they are stop words and are removed. On the other hand, high frequency terms could not be good indicators as the key terms to distinguish the relevant information in the given document. To solve this problem, researchers employ inverse document frequency (IDF) to diminish the weight of terms. IDF is a balanced way to offset unimportant terms with high frequency in the documents. Liu et al. (2008) reported that a word with a low IDF means that it occurs in many documents and is not topic indicative. Moreover, a term could usually appear more frequently in a document but rarely in other documents, which could significantly represent the topic. Hussey et al. (2012) emphasised that high weight (indicating importance) was achieved by having a high TF in the given document and a low occurrence in the remaining documents in the corpus. Therefore, IDF is a measurement of reduced weight for high frequency terms in remaining documents. The IDF formula is:

$$idf(t) = \log \frac{|D|}{|\{d : t \in d\}s|}$$

- D is the entire number of documents in the corpus
- d is the quantity of documents which contain the term t
- *idf(t)* is the inverse document frequency about the term t

TF-IDF weight is a statistical measurement to determine if a term is important to the document and uncommon to the corpus. It was first published in 1972 (Lott, 2012) and many new extraction applications still rely on the theory. The formula is:

$$tfidf(t,d) = tf(t) * idf(t)$$

TF-IDF weight assesses the importance of a word to a document in a collection and the word with higher TF-IDF scores is significant to the document and can summarise the document (Sparck Jones, 1972). It is widely used today in information retrieval, text summarisation and text mining. However, the disadvantage of TF-IDF is that it counts the frequency for a particular word, without considering any words that are similar to it in terms of semantic meaning. In addition, when the document is short, the TF-IDF may not be a reliable indicator of the importance of the word (Liu et al., 2009).

Table 4.1: TF-IDF scores with their corresponding words for the first eight questions

Is Rheumatoid Arthritis more common in men or women? [(0.63), (0.37),(0.25)] [u'woman', u'ra', u'men'],
Are there any DNMT3 proteins present in plants? [(0.71), (0.28), (0.20)] [u'novo', u'methyltransferase', u'drm2']
What is the most prominent sequence consensus for the polyadenylation site? [(0.36), (0.35), (0.30)] [u'aauaaa', u'polyadenylation', u'poly']
What is the function of the mammalian gene Irg1? [(0.55, (0.31), (0.24)] [u'irg1', u'gene', u'implantation']
Is thrombophilia related to increased risk of miscarriage? [(0.60), (0.42), (0.29] [u'pregnancy', u'thromboph ilia', u'woman']
Which extra thyroid tissues have thyrotropin (TSH) receptors? [(0.44), (0.27), (0.27)] [u'orbital', u'tshr', u'cd34']
What is the methyl donor of DNA (cytosine-5)-methyltransferases? [(0.35), (0.25), (0.24)] [u'methyl', u'donor', u'cytosine']
Super-SILAC is a method used in quantitative proteomics. What is the super-SILAC mix? [(0.49), (0.25), (0.24)] [u'silac', u'super', u'label']

In the first experiment in the BioASQ dataset, we apply the TF-IDF formula modified for our dataset, then rank TF-IDF scores from the highest to lowest scores. The TF-IDF scores and their corresponding words are shown in Table 4.1. These questions are retrieved from BioASQ dataset. It is a sample of the output set which has selected three terms, omitting decimal fractions smaller than 0.005 and counting all others. These are the first eight questions from the BioASQ dataset.

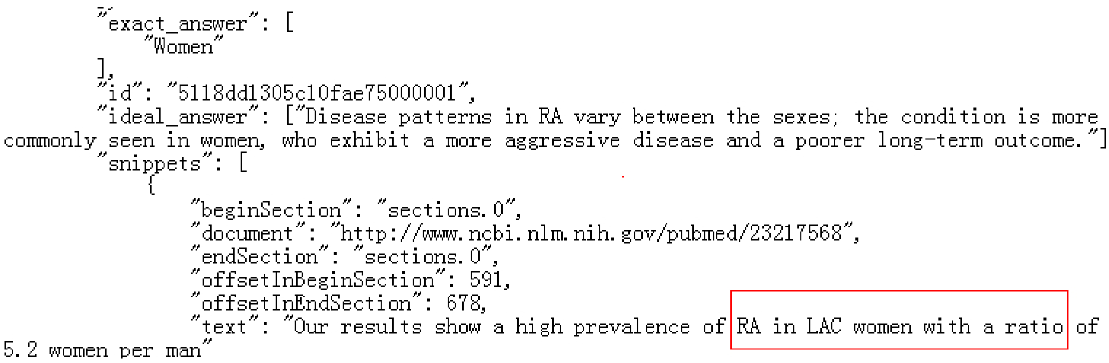
To ensure a reliable test, it is important to repeat the TF-IDF approach using a different number of keywords. Table 4.2 shows the top five terms listed using TF-IDF for the first, second and third questions. In our research, we use 3, 5, 8, 10 and 15 as the quantity of keywords set. It should be noted that synonyms still remain. For example, we cannot identify that the words "females" and "woman" have the same meaning. Due to time constraints, we are not able to investigate the use of biomedical thesaurus and other methods to identify synonymous terms.

TABLE 4.2: Sample of top five keywords by TF-IDF for the first three questions

Is Rheumatoid Arthritis more common in men or women? [u'women', u'ra', u'men', u'female', u'male']
Are there any DNMT3 proteins present in plants? [u'novo', u'methyltransferase', u'drm2', u'dna', u'methylation']
What is the most prominent sequence consensus for the polyadenylation site? [u'aauaaa', u'polyadenylation', u'poly', u'aataaa', u'signal']

4.3 Extraction approach 2 – Neighbourhood keywords extraction

The BioASQ dataset provides two types of concise answers: "ideal_answer" and "exact_answer". In Chapter 3, we presented the "ideal_answer" as the paragraph-sized summary and the part of "exact_answer" is symptom, disease and genes. As the concise answer, "exact_answer" contains the names of particular diseases or genes that are core information. Having core information, the "exact_answer" could help our project to highlight the neighbour words and even themselves. It is worth trying to explore the "exact_answer" in the text fragments. Therefore, the "exact_answer" can be the clue to explore the potential information by the structure of the sentence. Then our second experiment uses the "exact_answer" as the assigned central word to extract the neighbourhood words around "exact_answer" in the "text" of "snippets". Figure 4.1 shows the position of "exact_answer" and the neighbour words in the text fragment. Here we notice the "RA (Rheumatoid Arthritis)" is close to "women" and is extracted by the TF-IDF approach as well.



```
        "exact_answer": [
            "Women"
        ],
        "id": "5118dd1305c10fae75000001",
        "ideal_answer": ["Disease patterns in RA vary between the sexes; the condition is more
commonly seen in women, who exhibit a more aggressive disease and a poorer long-term outcome."],
        "snippets": [
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
                "endSection": "sections.0",
                "offsetInBeginSection": 591,
                "offsetInEndSection": 678,
                "text": "Our results show a high prevalence of RA in LAC women with a ratio of
5.2 women per man"
```

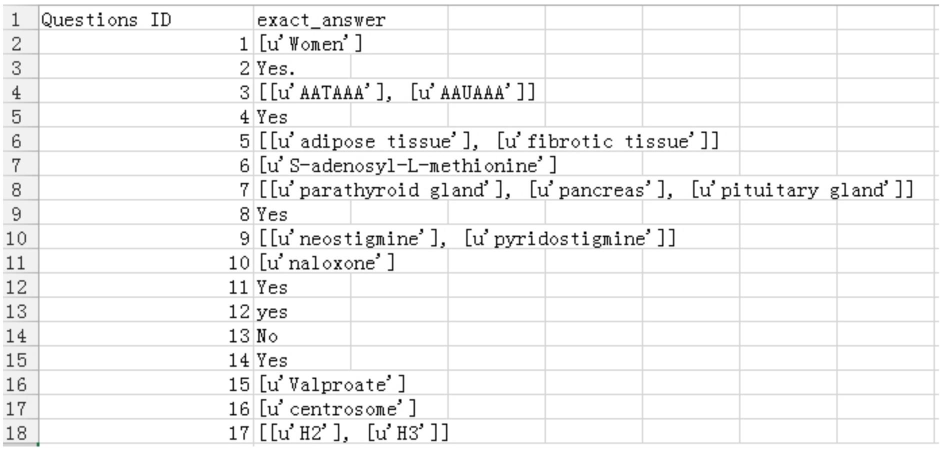
Figure 4.1: "Exact_answer" and neighbour words

The algorithm is given a specified term that is the "exact_answer", which is provided by participants or biomedical experts. The neighbourhood keyword extraction aims to find a few nearest words or phrases by the distance of the "exact_answer". This method could be a good way to explore terms because "exact_answer" is a indicator. Table 4.3 shows the extractd neighbourhood words with "exact_answer" for the first five questions.

Table 4.3: Sample of neighbourhood words with 'exact_answer' for the first five questions

[u'exhibit', u'prebiologic', u'significantly', u'prevalence', u'079', u'observed', u'manRA', u'67', u'seen', u'aggressive', u'total', u'presents', u'rated', u'ratio', u'Responses', u'common', u'per', u'1', u'prognosis', u'3', u'2', u'5', u'4', u'DAS28', u'9', u'clinical', u'RA', u'A', u'females', u'worse', u'082', u'men', u'125', u'despite', u'225', u'years', u'condition', u'women', u'15', u'14', u'progression', u'In', u'55', u'expression', u'LACThe', u'disease', u'commonly', u'n', u'patients', u'era', u'LAC', u'activity', u'found', u'higher', u'432']
u'yes.'
[u'required', u'sequence', u'polyadenylation', u'AATAAATwo', u'site', u'poly', u'appropriate', u'sequences', u'intact', u'results', u'two', u'UGUAAAFunctional', u'mutated', u'coded', u'motifs', u'AAUAAA', u'function', u'elements', u'G', u'spacing', u'Two', u'U', u'RNA', u'present', u'box', u'last', u'signal', u'AATAAA', u'hexanucleotide', u'consensus', u'demonstrate', u'formationthe']
u'yes'
[u'fibrocytes', u'lipolysis', u'TSH', u'These', u'lipolytic', u'tissues', u'adipose', u'factor', u'worked', u'induced', u'white']

In Chapter 3, the "exact_answer" has three types of responses: "yes" or "no" response for yes/no questions, a named entity for factoid questions, and a list of named entities for list questions (BioASQ, 2014b). In Table 4.3, there are two "yes" responses. Since entities and factoids are the clues to explore keywords, we extract all of the "exact_answers" and import them to a CSV file to check how many entities or factoids can be used.



1	Questions ID	exact_answer
2	1	[u'Women']
3	2	Yes.
4	3	[[u'AATAAA'], [u'AAUAAA']]
5	4	Yes
6	5	[[u'adipose tissue'], [u'fibrotic tissue']]
7	6	[u'S-adenosyl-L-methionine']
8	7	[[u'parathyroid gland'], [u'pancreas'], [u'pituitary gland']]
9	8	Yes
10	9	[[u'neostigmine'], [u'pyridostigmine']]
11	10	[u'naloxone']
12	11	Yes
13	12	yes
14	13	No
15	14	Yes
16	15	[u'Valproate']
17	16	[u'centrosome']
18	17	[[u'H2'], [u'H3']]

Figure 4.2: Sample of extracted "exact_answer"

Figure 4.2 presents the part of samples of "exact_answer". As some types of "exact_answer" are "yes" or "no", we cannot use them as the identifier to search the neighbourhood words and phrases. On the other hand, as discussed in Chapter 3, the "exact_answer" is not provided for every question unit. The use of "exact_answer" could not be enough to support this test. Both of the two factors would affect the evaluation results.

4.4 Extraction approach 3 – Noun phrases filter

Keywords are a set of significant words in an article that give a high-level description of the contents to readers and are useful to produce a short summary of an article (Lee & Kim, 2008). It is noticed that most of the keywords are nouns and verbs. This is the same situation in keyphrases. Manually identifying noun phrases is an extremely time-consuming and difficult task, which is nearly impossible to implement in multi-documents due to huge volumes. For the rapid extraction, we need the noun phrases patterns to automatically filter out matched phrases.

In this experiment we use noun phrases filter to identify keyphrases, because noun phrases filter has many advantages including a small set of precise and meaningful patterns, exact extraction, ease of implementing and improving. In contrast, the disadvantages are that it is typically not very robust and more patterns have to be added to explore certain types.

Here, we briefly introduce the regular expression in order to better understand the process.

,[],(): for grouping elements start from the innermost
— or, e.g., (Adj — Noun) means Adjective or Noun
+: one or more times the preceding element, e.g. Noun+ Noun refers to two or more nouns (key phrase extraction)
?: zero or once the preceding element

*: zero or more times the preceding element

Although, regular expression is very powerful for extraction and is introduced in Chapter 3, it is still not suitable for our testing dataset due to less semantic and syntax relating. Here, we borrow the regular expression layout to present the noun phrases filter. Since they are very common in the English language, we use the following three noun phrases patterns in the filter.

1. $Noun^{+}Noun$
2. $(Adj|Noun)^{+}Noun$
3. $((Adj|Noun)^{+}|((Adj|Noun)^{*}(NounPrep)?)(Adj|Noun)^{*})Noun$

As the first step, we use the patterns to extract phrases. We noticed every text fragment within "snippets" always has a hyperlink. Every hyperlink provides the source of the text fragment such as author(s), publication and date. The contents of webpages are mainly the abstract and conclusion of the article. Adjunct Professor Hanna Suominen at National Information and Communications Technology Australia suggested using the full abstracts to explore the noun phrases, instead of the text fragment in the BioASQ dataset.

Based on her opinion, we design the program to crawl the full length text in the webpages. Text in the webpages is from MEDLINE database. Hanna Suominen helped filter out matched phrases using the resources we collected. Table 4.4 shows the extracted keyphrases with their link webpages' ID. In this experiment, it is not possible to consider the importance of keyphrases. In the next section, we combine C/NC-values to select keyphrases.

Table 4.4: Sample of extracted keyphrases with noun phrases filter

10080180 nuclear lamina x-linked forms
10082816 fn stimulation, cortical excitability
10090061 rhesus monkeys, white matter
10103340 end of vancomycin therapy, end of vancomycin, rise in serum, serum vancomycin concentrations, peak vancomycin, vancomycin concentration, vancomycin therapy, infants with peak vancomycin serum, peak vancomycin serum concentrations, disease states, serum vancomycin, rise in serum creatinine, infants with peak vancomycin, vancomycin serum concentrations, peak vancomycin serum, peak vancomycin concentration, drug therapy, serum concentrations, group b, serum creatinine, vancomycin serum,

4.5 Extraction approach 4 – C-value/NC-value

The three terms extraction approaches implemented so far are either from a statistics model or linguistics model. For comprehensive testing of the BioASQ data, we further consider the frequency of occurrence with a linguistic feature model — C-value. The other consideration is how to identify the importance of phrases in previous experiments. C-value enhances the common statistical measure of frequency term extraction, making it sensitive to particular types of multi-word terms and the nested terms. Secondary, C-value gives extraction of term context words (words that tend to appear with terms) and the incorporation of information from term context words to the extraction of terms (Frantzi et al., 2000). C-value can automatically extract string (term) or string nested terms from the documents. The C-value formula is:

$$C - value(a) = \begin{cases} \log_2 |a| \cdot f(a) \\ \log_2 |a| \, (f(a) - \dfrac{1}{p(T_a)} \sum_{b \in T_a} f(b)) \end{cases}$$

The first formula is for an independent string and the second formula is for the string nested in a term, where

- a is the candidate string
- *f(a)* is the term frequency of occurrence in the corpus terms
- *Ta* is the set selected terms containing a
- *f(b))* is the total frequency of the term which contains a term
- *P(Ta)* is the number of terms

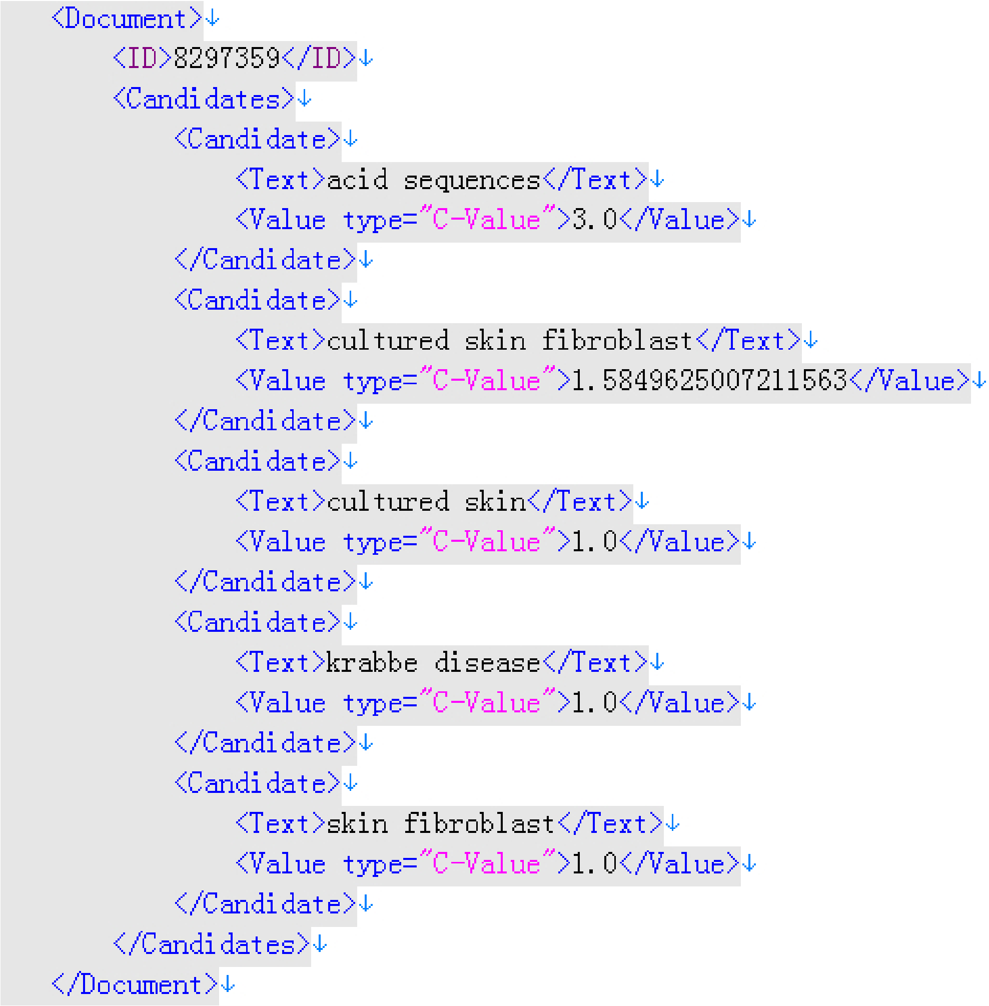
The C-value needs two parameters — the frequency of the candidate terms and the number of long candidate terms to calculate. In addition, NC-value is an improved algorithm of C-value. It incorporates context information into the C-value method for extraction of multi-word terms. The NC-value re-computes the scores without deleting and adding any terms which are returned by C-value. This approach adds the content of C-value to extract the multi-word terms. And it can assign the different weights to adjust the outputs:

- a is the candidate string
- *C(a)* is the set of distinct context words of a and b is a word from a
- $f_a(b)$ is the frequency of b as a term context word of a, weight(b) is the weight of b as a term context word (Frantzi et al., 2000).

$$NC - value(a) = 0.8C - value(a) + 0.2 \sum_{b \in c_a}^{f} a(b) weight(b)$$

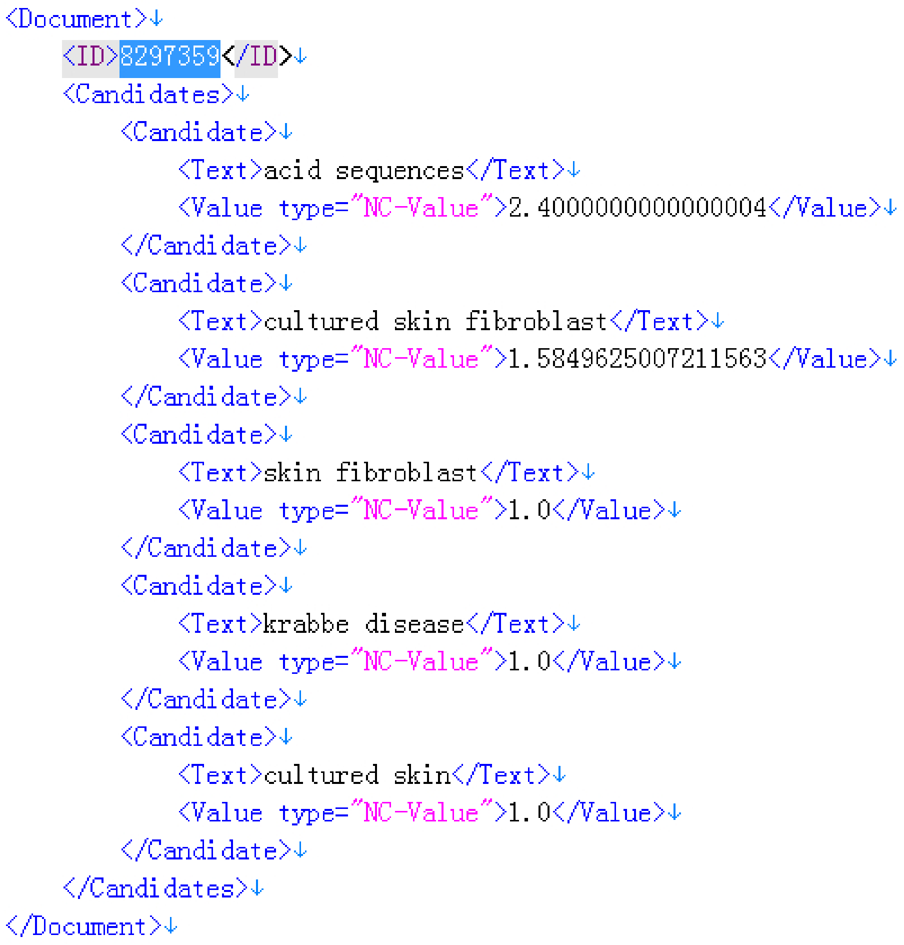
Here we used C/NC-values results in XML files developed by Hanna Suominen using C/NC-values tools. In our research project, we aim to extract terms directly. However, the information stored in XML files by C/NC-values is not easily accessible and it is necessary to transfer XML into an accessible format by our design. Here, we clean the outputs and remove the unnecessary tags and their values so that we can explore distilled information. Ultimately, we only use the keyphrases and their scores.

The fourth and fifth experiments use C/NC-values. The starting point of this work is to use the last experiment's output which are noun phrases collection and could be accepted by C/NC-values software. Based on C/NC-values results from Hanna Suominen's work, we filter out unnecessary parts successfully as shown in Figure 4.3. Figure 4.3 presents the candidate keyphrases with their C-value scores. We take the C-value score as the criteria to select the keyphrases. Figure 4.4 shows the NC-value scores with the same link ID of C-value. We highlight the first ID to distinguish between C-value and NC-value.



```
<Document>↓
    <ID>8297359</ID>↓
    <Candidates>↓
        <Candidate>↓
            <Text>acid sequences</Text>↓
            <Value type="C-Value">3.0</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>cultured skin fibroblast</Text>↓
            <Value type="C-Value">1.584962500721156З</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>cultured skin</Text>↓
            <Value type="C-Value">1.0</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>krabbe disease</Text>↓
            <Value type="C-Value">1.0</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>skin fibroblast</Text>↓
            <Value type="C-Value">1.0</Value>↓
        </Candidate>↓
    </Candidates>↓
</Document>↓
```

Figure 4.3: Keyphrases with their C-value scores



```
<Document>↓
    <ID>8297359</ID>↓
    <Candidates>↓
        <Candidate>↓
            <Text>acid sequences</Text>↓
            <Value type="NC-Value">2.4000000000000004</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>cultured skin fibroblast</Text>↓
            <Value type="NC-Value">1.5849625007211563</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>skin fibroblast</Text>↓
            <Value type="NC-Value">1.0</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>krabbe disease</Text>↓
            <Value type="NC-Value">1.0</Value>↓
        </Candidate>↓
        <Candidate>↓
            <Text>cultured skin</Text>↓
            <Value type="NC-Value">1.0</Value>↓
        </Candidate>↓
    </Candidates>↓
</Document>↓
```

Figure 4.4: Keyphrases with their NC-value scores

4.6 Summary

In this chapter, we have used selected extraction approaches to explore BioASQ dataset. These approaches we have implemented are without domain knowledge. The TF-IDF calculates the term frequency in one given document and decreased the weight of a term in the whole collection of documents. The neighbourhood keywords extraction uses the "exact_answer" as the indicative key information to help search the words near "exact_answer". The noun phrases filter is significant to generate phrases which are matched on the noun phrase patterns. Finally, we implement the C/NC-values built on linguistic and statistical characteristics. In noun phrases filter and C/NC-values approaches, we use the same phrases collection to test and compare the performance of C/NC-values.

One bottleneck is we cannot choose a certain number of keywords and keyphrases from neighbourhood keywords extraction and noun phrases filter approaches' outputs. The two terms extraction techniques do not attempt to justify the importance of words and phrases. For neighbourhood keywords extraction, the one improvement is to assign corresponding position weight to the words located near central terms and calculate the words in different positions. Due to a large number of noun phrases and selecting importance, we use C/NC-values to rank the noun phrases filter experiment's results.

These extraction approaches return sufficient results and we have collected keywords and keyphrases. Next Chapter, we evaluate these keywords and keyphrases whether or not they are key information of answers. We also investigate the extent which approach is the most effective for the BioASQ dataset.

5

Evaluation of five extraction approaches

5.1 Introduction

A number of extraction techniques were applied to the BioASQ dataset in Chapter 4. In this chapter, we evaluate the performance of these extraction approaches. The evaluation helps to better understand what measures can be used to assess value to actual users, and how to tailor their algorithms to meet users needs. The goal of our evaluation is to decide whether all selected extraction approaches can be used to answer biomedical questions. To measure the extraction approaches performance, we use precision, recall and F1-metric. Since evaluation is a critical issue, this chapter introduces evaluation frameworks. In Section 5.2 we explain the precision, recall and F1-metric. In Section 5.3 we present the evaluation framework and the measure of accuracy. In Section 5.4 we then show the evaluation of each type of experiment.

5.2 Evaluation metrics

5.3 Evaluation

We investigate the application of information retrieval evaluation. There are two general evaluating frameworks. The first method is applied by human annotation or human judgement. It needs domain expertise or specialists, which requires a huge amount of labour to label specific information. The second method is automatic evaluation which can save time and cost less. Our research project uses the second method to evaluate performances of all experiments.

The majority of systems compare their results with a "gold standard" in information retrieval. Due to the "ideal_answer" as the gold answer (reference answer) provided by biomedical experts, we take it as the gold standard to evaluate against the results. Extracted terms are evaluated against the full length of the "ideal answer" — gold response in the BioASQ dataset. Our evaluating goal is to ensure these proposed extraction techniques are suitable to find the interesting keywords and keyphrases in the "ideal_answer", so that a subsequent abstractive summariser would eventually generate the "ideal_answers" in the future.

In addition, the extracted terms can be a complete match, partial match, or no match to standards. In a complete match, the extracted terms totally appear in comparative text. In a partial match, in some way, the extracted terms are not totally included in the standard text. In a no match, the terms which are obtained are not identical and do not intersect in evaluating text. With the reference answer, it is possible that extracted terms are close enough to be complete matches and partial matches in the evaluation text. Our aim is to maximise the number of terms which are relevant words or phrases in the gold standard. In contrast, we may obtain incomplete information which is not significant, rather than return nothing.

5.3.1 Precision and recall

Given the truth of dataset D, we assign the correct items retrieved by the positive class called true positives (tp), where the correctly items are in dataset D; let false positives (fp) are the number of items are classified incorrectly and labeled not in D; As false negatives (fn), are items belonging to the positive dataset D but should have been retrieved. True negatives (tn) are incorrect and not retrieved.

Applying evaluating application, there are two evaluation metrics: precision and recall. Precision is the percentage of true positives in the retrieved results. True positives are retrieved and relevant. False positives are retrieved but non-relevant. Here n is equal to the total number of retrieved items: (tp + fp). The precision formula is:

$$precision = \frac{t_p}{t_p + f_p} = \frac{t_p}{n}$$

Recall is the percentage of true positives in the truth dataset, and fn is false negative, which is relevant but non-retrieved. The formula is:

$$recall = \frac{t_p}{t_p + f_n}$$

Precision and recall affect each other and there is an inverse relationship between them. For example, where the experiment has obtained a higher precision, then the related recall will be lower. It is, therefore, necessary to ensure that they are balanced. It should be noted that an increase in precision or recall does not necessarily correlate with user success in the searching task (Hersh et al., 2002).

In the case of our evaluation, precision is the ratio of the number of words from the "ideal_answer" retrieved to the total number of irrelevant and relevant words retrieved. It is usually expressed as a percentage. Recall is the ratio of the number of words from the "ideal_answer" retrieved to the total number of words in the "ideal_answer" that are in common with the text. It is expressed as a percentage as well. Thus, we propose effective evaluation paradigms in this research project. The paradigms are presented under the following classification:

- Precision = total_matches/ total_extracted_terms
- Recall = total_matches / total_ideal_answer

5.3.2 F1-metric

Manning & Schütze (1999) reported there are three advantages of using precision and recall: 1) accuracy figures are not very sensitive to the small, but interesting numbers of tp, fp, and fn whereas precision and recall are; 2) other things being equal, the F-metric prefers results with more true positives, whereas accuracy is sensitive only to the number of errors; 3) using precision and recall, one can give a different cost to missing target items versus selecting junk (noisy data). It can be convenient to combine precision and recall into a single measurement for the overall performance. One way to do this is the F-metric (F-measure). F-metric is the combination of precision and recall. Below is the formula for the classic F-metric and β is the argument.

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Simply, we assign β equivalent to 1 and the F-metric becomes the F1-metric.

$$F_1 = \frac{2 * PR}{P + R}$$

5.4 Results of evaluation

The evaluating task is to identify whether extracted terms exist in the "ideal_answer". If extracted terms are matched with the terms in the "deal_answer", these results are defined as "relevant", and the rest of text is defined as "not relevant". The process of evaluation is to match how many common terms there are between the set of extracted terms and "ideal answer". The successful matched terms are true positives and the rest are false positives which are contained in the set of extracted terms rather than in the "ideal_ answer", or false negatives which are not extracted in the set of terms but exist in the "ideal_answer".

For instance, the fifth question in the BioASQ dataset is "is thrombophilia related to increased risk of miscarriage?" and the related "ideal_answer" is "Thrombophilia has been found to be considerably more common in women with pregnancy-associated complications in comparison with the general population, and most frequently in conjunction with venous thromboembolism during pregnancy and the postpartum period. In particular there is an increased risk of pregnancy-related venous thrombosis in carriers of severe inherited thrombophilia. When counseling white women with a history of preeclampsia, screening for thrombophilia can be useful for preconceptional counseling and pregnancy management." The keywords extraction using TF-IDF with the highest three score words are [u'pregnancy', u'thrombophilia', u'women'] for this question. The three keywords exist in the "ideal_answer", which are considered as answers. Therefore, the evaluation of this question returns a precision of 1 with three keywords.

5.4.1 Evaluation of approach 1 – TF-IDF

In Chapter 4, these keywords and keyphrases are raw materials from original text. These raw materials are preprocessed and them sent to be evaluated. In every experiment we set different terms number, and the purpose is to investigate how the terms number influences keyword and keyphrase extraction performance. The number of terms in the set are 3, 5, 8, 10, and 15. We use the average precision, recall and F1-metric to evaluate the extracted terms depending on TF-IDF scores.

TF-IDF is the first approach we applied. Due to the consideration of potential knowledge, in the first run we keep all texts and semantic structure in "ideal answer", and in the second run we preprocess the raw text of "ideal_answer". There is a statistically significant difference in precision and recall between the two runs. Table 5.1 shows the average scores of precision, recall and F1-metric in the first run. Then we preprocess the "ideal_ answer" in order to remove the noise data, by removing these uninteresting words and symbols. Table 5.2 shows second run and a similar trend in the upper bound of the average precision, recall and F1-metric. For example, among the three terms set, the first run shows the average precision of 42.9% and recall of 4.1%. In the second run, the average precision is 70.1% and recall is 11.8%. We consider the second run as the final result because preprocessing removes uninteresting words and reduces noisy data, which is valid. In the following experiments, we totally preprocess the "ideal answer". When choosing 3 as the number of terms, the highest average precision is 70.1%. When choosing 15, we obtain the highest recall of 29.0% and F1-metric of 29.67%. When the number of terms increases, the average precision decreases. In contrast, the average recall and F1-metric increases.

Table 5.1: Evaluation of approach 1- TF-IDF without preprocessing

Number of keywords	Average Precision %	Average Recall %	Average F1-metric %
3	42.9	4.1	6.4
5	39.3	5.8	8.6
8	36.7	8.0	11.3
10	34.3	9.1	12.3
15	29.8	11.3	13.9

Table 5.2: Evaluation of TF-IDF performance 2

Number of keywords	Average Precision %	Average Recall %	Average F1-metric %
3	70.1	11.8	17.8
5	63.2	16.3	22.4
8	55.4	21.3	23.3
10	51.4	23.9	27.8
15	44.1	29.0	29.7

5.4.2 Evaluation of approach 2 – Neighbourhood keywords extraction

The assigned terms ("exact_answer") may have clues or influence the neighbour words mentioned in Chapter 4. Our study presents the assigned terms with k nearest neighbour words, which may become the keyword and keyphrase. Around annotated terms, they have possible associations in a sentence. Our approach is simple: we consider the distance from the "exact_answer" that is central terms. In order to investigate how the distance of central terms influences the keyword extraction performance, the distance with is set 3, 5, 8, 10 and 15. It is should be noted that we design this extraction approach without using the fixed quantity of terms.

There are 310 question units presented in Chapter 3 and there is one question unit without "snippets" which cannot be used to extract terms (see Appendix A.3). The reason may be that BioASQ had not collected the data or there is missing data. Then the dataset has 309 questions to be explored: 84 questions do not provide the "exact_answer" and in 75 questions the "exact_answer" is "yes", "no", "Yes", and "No" in our remaining testing dataset. Therefore, nearly half of the questions cannot use "exact_answer"as the specified terms to search neighbour words.

The second approach is neighbourhood terms extraction. The "exact_answer" is the concise answer to identify the question and "exact_answer" is assigned as the central terms to extract the closest words in text fragments. We count these "exact_answers" themselves as extracted terms as well. Using annotated terms, we suspect the neighbourhood terms extraction might perform better than the TF-IDF because it uses additional information about the answer. In Table 5.3, the highest precision is 2.9% when taking 3 terms as the number of keywords. The highest F1-metric is 11.2% when taking 8 terms. The approach of TF-IDF has shown stronger performance than the neighbourhood terms extraction.

TABLE 5.3: Evaluation of approach 2 – neighborhood keywords extraction

Distance	Average Precision %	Average Recall %	Average F1-metric %
3	2.9	1.6	1.0
5	2.8	2.4	1.2
8	2.7	2.8	1.2
10	2.6	2.9	1.2
15	2.5	3.1	1.1

First, Table 5.3 shows the evaluation of all questions. Because the “exact answer” is not provided in every question unit, the empty values or the “yes/no” types in “exact answer” cannot be used in this experiment and cannot be evaluated either. The default values of precision and recall are set at 0. There is a significant decrease in performance compared to evaluation of TF-IDF, as there are not enough questions where we can extract terms. Second, part of “exact_answer” is long phrase such as, “*Administration of 7-nitroindazole (neuronal nitric oxide synthase inhibitor)*” in the question “*List some ways to reverse Tau hyperphosphorylation in Tauopathies?*”. Part of such types of “exact_answer” do not exist in the “text” in BioASQ dataset. This could be another reason affecting the extraction performance, because very few neighbourhood keywords cannot provide sufficient information and they may introduce noisy data. The performance of neighbourhood keyword extraction is not effective and in future we can improve its efficiency.

5.4.3 Evaluation of approach 3 – Noun phrases filter

The third experiment of extracting terms is the noun phrases filter recognition approach. Extracting phrases with these patterns means that other unidentied types of phrases are lost, that is, a number of unidentified phrases are not extracted in this test. Based on Chapter 4, we already demonstrate the very common three noun phrases patterns in English.

In this experiment using noun phrases filter, we explore more text other than only fragments, which are full length abstracts. We suppose the evaluation of noun phrases filter is stronger than the evaluation of the previous experiment. Table 5.4 shows the average precision, recall and F1-metric on the testing data. The evaluation result has shown it is worse than the evaluation of TF-IDF, but much better than neighbourhood keywords extraction. We ran the experiment five times and every time we selected a different number of candidate keyphrases. The number of candidates keyphrases are 3, 5, 8, 10, 15. By precision, recall and F1-metric, the results are still at a higher precision corresponding to a lower recall. But the average precision is 37.5% and recall is 4.2% when choosing 3 keyphrases, which is still lower than TF-IDF. The best average F1-metric is 11.3% when the number of keyphrases is 15.

Table 5.4: Evaluation of approach 3 – Noun phrases filter

Noun phrases	Average Precision %	Average Recall %	Average F1-metric %
3	37.5	4.2	6.8
5	31.5	5.9	8.7
8	26.2	8.0	10.3
10	23.7	8.8	10.7
15	19.6	10.7	11.3

The performance of noun phrases filter for the BioASQ dataset is worse than the TF-IDF method. The lower precision and recall is reasonable, since at this stage we only select noun phrases. In addition, the lower recall highlights that many noun phrases we explored do not exist in the "ideal_answer", due to the full length text of webpages. Finally, every keyphrase is divided into single words to match the "ideal_answer", which could be a limitation to the evaluation.

5.4.4 C/NC-values

Because C/NC-values approaches are powerful, we expand the source of testing text with the online abstract which is described in Chapter 4. That is, this approach incorporates more text which can explore more phrases. We do not set the length of candidate terms.

As with previous experiments and evaluations, we use precision, recall and F1-metric to compare them with the number of different terms. We need a set of terms, and still choose the number 3, 5, 8, 10 and 15. For NC-value evaluation, we choose the same number of terms as in the evaluation of C-value. Table 5.5 and Table 5.6 show the evaluation of C-value and NC-value: the evaluation of NC-value is slightly worse than C-value.

The last evaluation contains the fourth and fifth experiments. With C/NC-values, performance in evaluation is extremely poor. As C-value, the highest average of precision is 1.9% and the recall is 0.5%, when extracting 3 keyphrases. The best performance of F1-metric is 1.0% when the number of keyphrases was increased to 15. For NC-value, the average of precision and recall is lower than the C-value. The highest F1-metric is 1.0%. The C/NC-values performs poorly for the BioASQ dataset.

There are several explanations for the poor performance. Firstly, these "text" in BioASQ dataset are relevant answers which could be used to identify the question and are highly

Table 5.5: Evaluation of C-value

C-value terms	Average Precision %	Average Recall %	Average F1-metric %
3	1.9	0.5	0.7
5	1.7	0.7	0.9
8	1.4	0.9	1.0
10	1.4	1.1	1.0
15	1.1	1.4	1.0

Table 5.6: Evaluation of NC-value

NC-value terms	Average Precision %	Average Recall %	Average F1-metric %
3	1.8	0.5	0.1
5	1.6	0.6	0.8
8	1.4	0.9	0.9
10	1.3	1.0	1.0
15	1.1	1.4	1.0

correlated. We expanded the scope of resources beyond the “text” in the well-structured BioASQ dataset, which adds more non-topical phrases that are not correlated. Secondly, nested terms appeared more often in other longer terms, and nested terms may not be independent. These longer terms are not evaluated in the “ideal answer”. Thirdly, these extracted terms may not be the real terms, which are presented in Table 4.4. Domain experts are needed to interpret and judge whether or not they are real terms. So far, we have no automatic indication for them. In Figure 4.3, it is clear that the extracted terms have the same C-value scores, like “ cultures skin” and “krabbe disease”. The flaw of this experiment is that it does not set priorities for these terms which have the same C/NC-values. It should be noted that the TF-IDF experiment uses the “text” in the BioASQ dataset to calculate scores whereas the C/NC-values experiment uses the full length text in abstracts to compute scores. This is the reason more extracted terms have the same C/NC-values than extracted terms have the same TF-IDF scores.

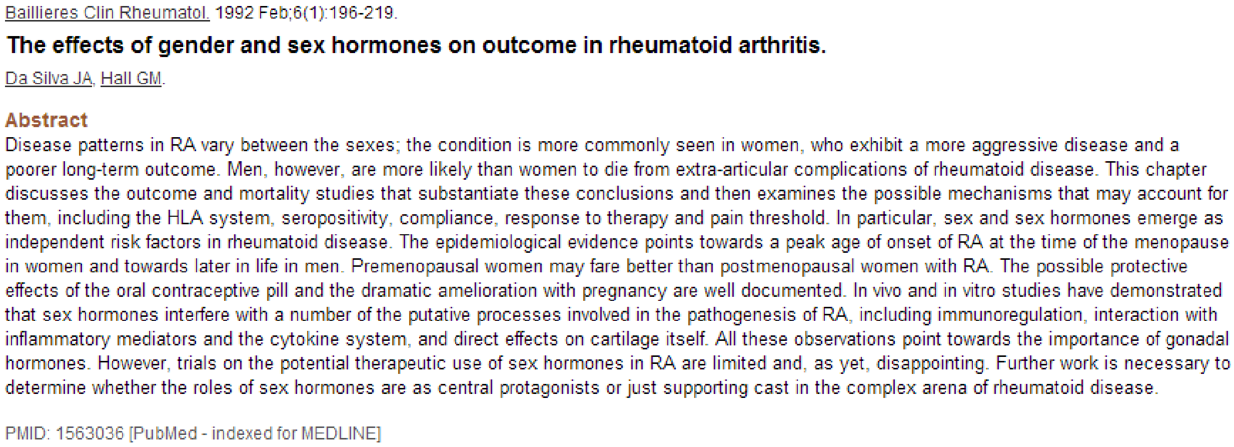
Based on Table 5.5 and Table 5.6, there is poor performances of C/NC-values. We further combine TF-IDF to rank these extracted keyphrases compared to C/NC-values. This comparison is to select the importance of extracting keyphrases, due to a large amount of noun phrases data in the third experiment. Using TF-IDF is the most common selection approach that is based on collection statistics. Table 5.7 presents the evaluation of performance using TF-IDF to rank extracted keyphrases, which further investigates the influence of TF-IDF on noun phrases. The evaluation result is stronger than evaluations of C/NC-values.

Table 5.7: Evaluation of noun phrases filter with TF-IDF

NC-value terms	Average Precision %	Average Recall %	Average F1-metric %
3	43.0	6.1	9.7
5	36.1	8.5	12.0
8	30.5	11.4	14.1
10	27.4	12.7	14.6
15	22.6	15.3	15.2

5.5 Problems

There are several limitations to the evaluation. During processing keywords, it should be noted that there was one question without snippets. Missing information can also lead to somewhat misleading conclusions (Blaschke et al., 1999). The second problem faced is possible lack of good evidence for the evaluation, by using retrieved text fragments in the BioASQ data. As many published research papers and scientific literature are not available as full-length texts due to copyright, text fragments are "text" in "snippets" which may not be comprehensive. For example, Figure 5.1 presents the abstract of an article in MEDLINE. The text fragments in BioASQ dataset are most from such abstracts and potential knowledge may not be contained in the text fragments. Thirdly, the comparison between "ideal_answer" and extracted terms highlights the performance of selected approaches. But the performance of some extraction approaches tested is very poor due to very lower F1-metric. Since matched terms corresponding to the question may become answers, these extraction approaches with lower F1-metric are not useful in answering biomedical questions.



Baillieres Clin Rheumatol. 1992 Feb;6(1):196-219.

The effects of gender and sex hormones on outcome in rheumatoid arthritis.

Da Silva JA, Hall GM.

Abstract

Disease patterns in RA vary between the sexes; the condition is more commonly seen in women, who exhibit a more aggressive disease and a poorer long-term outcome. Men, however, are more likely than women to die from extra-articular complications of rheumatoid disease. This chapter discusses the outcome and mortality studies that substantiate these conclusions and then examines the possible mechanisms that may account for them, including the HLA system, seropositivity, compliance, response to therapy and pain threshold. In particular, sex and sex hormones emerge as independent risk factors in rheumatoid disease. The epidemiological evidence points towards a peak age of onset of RA at the time of the menopause in women and towards later in life in men. Premenopausal women may fare better than postmenopausal women with RA. The possible protective effects of the oral contraceptive pill and the dramatic amelioration with pregnancy are well documented. In vivo and in vitro studies have demonstrated that sex hormones interfere with a number of the putative processes involved in the pathogenesis of RA, including immunoregulation, interaction with inflammatory mediators and the cytokine system, and direct effects on cartilage itself. All these observations point towards the importance of gonadal hormones. However, trials on the potential therapeutic use of sex hormones in RA are limited and, as yet, disappointing. Further work is necessary to determine whether the roles of sex hormones are as central protagonists or just supporting cast in the complex arena of rheumatoid disease.

PMID: 1563036 [PubMed - indexed for MEDLINE]

Figure 5.1: Example of abstract in MEDLINE
Source : http://www.ncbi.nlm.nih.gov/pubmed/1563036

5.6 Summary

In this chapter we have described the design of evaluations and compared different automatic terms extraction approaches using “ideal_answer” in the BioASQ dataset. A number of terms could be the answers of the specific biomedical questions in the text fragments. The successful matched terms could be such answers in comparison to the “ideal_ answer”. For our evaluation, we used precision and recall to measure candidate terms. TF-IDF outperformed the best and has the highest F1-metric scores. Potential improvements to the evaluation of extraction methods are discussed in the next chapter.

6

Results and Discussion

6.1 Introduction

This project assesses the ability to answer biomedical questions by using keywords and keyphrases extraction techniques. Here, we have shown keywords and keyphrases as the core concepts of abstractive summarisation, where medical practitioners like quick, short and easy answers. In Chapter 5 we presented the evaluation framework and evaluations of five approaches. In this chapter, we compare the five categories of extraction approaches and discuss future work. Section 6.2 provides the overview of evaluation for the selected approaches, Section 6.3 gives achievements of the research project and Section 6.4 presents the future research directions.

6.2 Comparison of approaches for terms extraction

6.2.1 Analysis

In this research project, we have developed algorithms to extract terms from text fragments of BioASQ. Our experimental approaches used TF-IDF, neighbourhood keywords extraction, noun phrases filter and C/NC-values. And we have evaluated whether these algorithms are appropriate for biomedical questions in the BioASQ dataset. Extracted terms are evaluated using the "ideal_answer" that is the validation set. In addition, these terms that could be used as the starting point for abstractive summarisation may be of great value for the purpose of answering biomedical questions. The evaluation showed using TF-IDF outperforms the other extraction approaches.

In our research the use of selected extraction approaches is sufficient to generate a range of results without further to implement syntax and semantic analysis. But it still needs domain experts to help justify and judge some outputs because there are fewer semantic relationships. For instance, we have obtained the terms “men” and “women” using TF-IDF in the first question. The “ideal_answer” has shown “the condition is more commonly seen in women than men.” As we can see, “men” and “women” are common words appearing in the extracted terms set and “ideal_answer”. Keyword extraction approaches cannot present semantic and syntactic structures.

For a query-based summarisation system in the medical domain, the most important factors are the diagnosis, the description of disease and the clinical treatment. However, extraction techniques for query-based summarisation systems are to explore true information, but cannot distinguish the quality of this information (Blaschke et al., 1999). A further drawback, we use common stop words list to filter out unnecessary words, which could not be precise. This process is not effective compared to using the biomedical stop words list.

6.2.2 Conclusion

TF-IDF and noun phrases filter have shown better performance for term extractions and TF-IDF achieved the highest F1-metric scores. Since short testing texts are relevant passages, they reflect the same topic in every question unit. The important keywords and keyphrases in short texts are repeated, which is not surprising. Based on our reported evaluation of performances, the research contributes to knowledge that TF-IDF is the best term extraction technique that can be used to answer biomedical questions. It is suggested that other selected extraction approaches do not compare favourably. On the other hand, by increasing the number of keywords and keyphrases, the evaluation obtained higher scores for all experiments except neighbourhood keywords extraction. When choosing eight terms, the neighbourhood keywords extraction achieved the highest F1-metric of 1.2%.

To present the overall performance of the proposed term extraction approaches, we select the top five terms as a baseline, since that is the number of terms human annotators use as a guideline (Liu et al., 2009). As quantity of selected keywords and keyphrases increase, the relevant F1-metric is increasing as well. In this research project, the 15 terms have achieved the highest F1-metric scoress. Table 6.1 shows the average F1-metric of every approach using top 5 and 15 terms, including noun phrases filter with TF-IDF.

Table 6.1: Average F1-metric of results from top 5 and 15 terms

Average F1-metric % terms number Extraction approaches	5 terms	15 terms
TF-IDF	22.4	29.7
Neighbourhood keywords extraction	1.2	1.1
Noun phrases filter	8.7	11.3
Noun phrases filter with TF-IDF	12.0	15.2
C-value	0.9	1.0
NC-value	0.8	1.1

6.3 Summary of achievements

In this research project, we have used five terms extraction approaches to explore the BioASQ dataset. The TF-IDF approach achieves the highest F1-metric scores and is the best performance in the evaluation. The results of TF-IDF are applicable and most appropriate to identify parts of the answer. TF-IDF can be used in answering biomedical questions and can explore evidence and these extracted terms can be used as evidence in medical practice. Noun phrases filter generates sufficient keyphrases as well but the performance of noun phrases filter is not as good as TF-IDF. Also this research project contributes abstractive summarisation that is just a place to start. Some results are not as efficient as expected, especially the neighbourhood keywords extraction. But it provides a path for the co-occurrence between terms and the query, such as "exact_answer" and the question in the BioASQ dataset.

6.4 Future research directions

In this work, we explore using extraction techniques to extract relevant terms as the answers to solve biomedical questions, which are potential knowledge. This work also generates several possible future research directions. For some reasons, the testing dataset is not able to provide the "snippets" or "exact answer" in every question unit. Thus, one shortcoming of the research project is that outputs are not fully evaluated. The immediately future work is to determine the importance of keywords and keyphrases which have the same TF-IDF scores or C/NC-values. Adding more patterns is worth considering, such as more complex noun phrase patterns or other types of phrase patterns. More patterns mean more varieties of phrases can be extracted and could achieve better results. Another future direction is to explore keyword and keyphrase ranking algorithms, which can be integrated into the testing framework.

Even well-formed medical texts have different characteristics from general domain texts, and require tailored solutions (Holzinger et al., 2007). Due to the complexity of natural language itself, future work can explore terms by using a thesaurus and abbreviations. The terms are usually restricted into a domain-specic thesaurus, which could be more precise for applications. Medical terminology should be taken into account as well, to exploit existing medical resources for specific diseases or biological threats (Afantenos et al., 2005). This is a direction in future work.

A

Appendix

A.1 Stop words list

Table A.1: Stop word list

a's	able	about	above	according	accordingly
across	actually	after	afterwards	again	against
ain't	all	allow	allows	almost	alone
along	already	also	although	always	am
among	amongst	an	and	another	any
anybody	anyhow	anyone	anything	anyway	anyways
anywhere	apart	appear	appreciate	appropriate	are
aren't	around	as	aside	ask	asking
associated	at	available	away	awfully	be
became	because	become	becomes	becoming	been
before	beforehand	behind	being	believe	below
beside	besides	best	better	between	beyond
both	brief	but	by	c'mon	c's
came	can	can't	cannot	cant	cause
causes	certain	certainly	changes	clearly	co

Table A.2: Stop word list part 2

com	come	comes	concerning	consequently	consider
considering	contain	containing	contains	corresponding	could
couldn’t	course	currently	definitely	described	despite
did	didn’t	different	do	does	doesn’t
doing	don’t	done	down	downwards	during
each	edu	eg	eight	either	else
elsewhere	enough	entirely	especially	et	etc
even	ever	every	everybody	everyone	everything
everywhere	ex	exactly	example	except	far
few	fifth	first	five	followed	following
follows	for	former	formerly	forth	four
from	further	furthermore	get	gets	getting
given	gives	go	goes	going	gone
got	gotten	greetings	had	hadn’t	happens
hardly	has	hasn’t	have	haven’t	having
he	he’s	hello	help	hence	her
here	here’s	hereafter	hereby	herein	hereupon
hers	herself	hi	him	himself	his
hither	hopefully	how	howbeit	however	i’d
i’ll	i’m	i’ve	ie	if	ignored
immediate	in	inasmuch	inc	indeed	indicate
indicated	indicates	inner	insofar	instead	into
inward	is	isn’t	it	it’d	it’ll
it’s	its	itself	just	keep	keeps
kept	know	known	knows	last	lately
later	latter	latterly	least	less	lest
let	let’s	like	liked	likely	little
look	looking	looks	ltd	mainly	many
may	maybe	me	mean	meanwhile	merely
might	more	moreover	most	mostly	much
must	my	myself	name	namely	nd
near	nearly	necessary	need	needs	neither
never	nevertheless	new	next	nine	no
nobody	non	none	noone	nor	normally
not	nothing	novel	now	nowhere	obviously
of	off	often	oh	ok	okay
old	on	once	one	ones	only
onto	or	other	others	otherwise	ought
our	ours	ourselves	out	outside	over
overall	own	particular	particularly	per	perhaps

Table A.3: Stop word list part 3

placed	please	plus	possible	presumably	probably
provides	que	quite	qv	rather	rd
re	really	reasonably	regarding	regardless	regards
relatively	respectively	right	said	same	saw
say	saying	says	second	secondly	see
seeing	seem	seemed	seeming	seems	seen
self	selves	sensible	sent	serious	seriously
seven	several	shall	she	should	shouldn't
since	six	so	some	somebody	somehow
someone	something	sometime	sometimes	somewhat	somewhere
soon	sorry	specified	specify	specifying	still
sub	such	sup	sure	t's	take
taken	tell	tends	th	than	thank
thanks	thanx	that	that's	thats	the
their	theirs	them	themselves	then	thence
there	there's	thereafter	thereby	therefore	therein
theres	thereupon	these	they	they'd	they'll
they're	they've	think	third	this	thorough
thoroughly	those	though	three	through	throughout
thru	thus	to	together	too	took
toward	towards	tried	tries	truly	try
trying	twice	two	un	under	unfortunately
unless	unlikely	until	unto	up	upon
us	use	used	useful	uses	using
usually	value	various	very	via	viz
vs	want	wants	was	wasn't	way
we	we'd	we'll	we're	we've	welcome
well	went	were	weren't	what	what's
whatever	when	whence	whenever	where	where's
whereafter	whereas	whereby	wherein	whereupon	wherever
whether	which	while	whither	who	who's
whoever	whole	whom	whose	why	will
willing	wish	with	within	without	won't
wonder	would	wouldn't	yes	yet	you
you'd	you'll	you're	you've	your	yours
yourself	yourselves	zero			

A.2 Fragment of BioASQ dataset

This part is the first question which is complete.

```
"questions": [
    {
        "body": "Is Rheumatoid Arthritis more common in men or women?",
        "concepts": [
   "http://www.disease-ontology.org/api/metadata/DOID:7148",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exac
    t=Find+Exact+Term&term=D001171",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact
   =Find+Exact+Term&term=D012217",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact
   =Find+Exact+Term&term=D013167",
   "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exac
   t=Find+Exact+Term&term=D015535"
        ],
        "documents": [
            "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
            "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
            "http://www.ncbi.nlm.nih.gov/pubmed/21340496",
            "http://www.ncbi.nlm.nih.gov/pubmed/20889597",
            "http://www.ncbi.nlm.nih.gov/pubmed/20810033",
            "http://www.ncbi.nlm.nih.gov/pubmed/19158113",
            "http://www.ncbi.nlm.nih.gov/pubmed/18759162",
            "http://www.ncbi.nlm.nih.gov/pubmed/17965425",
            "http://www.ncbi.nlm.nih.gov/pubmed/16418123",
            "http://www.ncbi.nlm.nih.gov/pubmed/15083883",
            "http://www.ncbi.nlm.nih.gov/pubmed/12723987",
            "http://www.ncbi.nlm.nih.gov/pubmed/1563036"
        ],
        "exact_answer": [
            "Women"
        ],
        "id": "5118dd1305c10fae75000001",
        "ideal_answer": ["Disease patterns in RA vary between the sexes;
        the condition is more commonly seen in women,who exhibit a more
        aggressive disease and a poorer long-term outcome."],
        "snippets": [
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
            "endSection": "sections.0",
            "offsetInBeginSection": 591,
            "offsetInEndSection": 678,
            "text": "Our results show a high prevalence of
             RA in LAC women with a ratio of 5.2 women per man"
        },
        {
            "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
            "endSection": "sections.0",
            "offsetInBeginSection": 1140,
            "offsetInEndSection": 1394,
            "text": "RA in LAC women is not only more common but presents
            with some clinical characteristics that differ from
            RA presentation in men. Some of those characteristics
            could explain the high rates of disability and worse prognosis
            observed in women with RA in LAC"
        },
        {
            "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
            "endSection": "sections.0",
            "offsetInBeginSection": 559,
            "offsetInEndSection": 718,
            "text": "The expression and clinical course of RA are
             influenced by gender. In developed countries the prevalence
             of RA is 0,5 to 1.0%, with a male:female ratio of 1:3."
        },
        {
            "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
            "endSection": "sections.0",
            "offsetInBeginSection": 897,
            "offsetInEndSection": 1031,
            "text": "women were found to have higher disease activity
            scores, more pain and greater loss of function, both
             in early and established disease"
        },
        {
            "beginSection": "sections.0",
            "document": "http://www.ncbi.nlm.nih.gov/pubmed/21340496",
            "endSection": "sections.0",
            "offsetInBeginSection": 993,
            "offsetInEndSection": 1062,
        "text": "Intense anti-CCP2 reaction was 19.8-fold
        higher in females vs. males,"
    },
    {
        "beginSection": "sections.0",
        "document": "http://www.ncbi.nlm.nih.gov/pubmed/20889597",
        "endSection": "sections.0",
        "offsetInBeginSection": 911,
        "offsetInEndSection": 944,
        "text": " men (n = 67) and women (n = 225)"
    },
    {
        "beginSection": "sections.0",
        "document": "http://www.ncbi.nlm.nih.gov/pubmed/20889597",
        "endSection": "sections.0",
        "offsetInBeginSection": 1808,
        "offsetInEndSection": 1943,
        "text": " Responses to treatment over time were better among
         men in this prebiologic era; women had worse progression
         despite similar treatment."
    },
    {
        "beginSection": "sections.0",
        "document": "http://www.ncbi.nlm.nih.gov/pubmed/20810033",
        "endSection": "sections.0",
        "offsetInBeginSection": 1550,
        "offsetInEndSection": 1629,
        "text": "BMI appears to be associated with RA disease activity
        in women, but not in men."
    },
    {
        "beginSection": "sections.0",
        "document": "http://www.ncbi.nlm.nih.gov/pubmed/20810033",
        "endSection": "sections.0",
        "offsetInBeginSection": 729,
        "offsetInEndSection": 785,
        "text": "A total of 5,161 RA patients (4,082 women and 1,079 men)"
    },
    {
        "beginSection": "sections.0",
        "document": "http://www.ncbi.nlm.nih.gov/pubmed/19158113",
        "endSection": "sections.0",
        "offsetInBeginSection": 561,
        "offsetInEndSection": 744,
                "text": "In women the DAS28 was significantly higher
                 than in men due to higher scores for general health
                 and tender joints. Likewise, HAQ and VAS pain were
                 rated significantly higher in women."
            },
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/18759162",
                "endSection": "sections.0",
                "offsetInBeginSection": 263,
                "offsetInEndSection": 285,
                "text": "432 females, 125 males"
            },
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/17965425",
                "endSection": "sections.0",
                "offsetInBeginSection": 862,
                "offsetInEndSection": 1017,
                "text": "ESR significantly increased with age, independent
                of other variables of disease activity. This increase was
                more pronounced in male than in female patients"
            },
            {
                "beginSection": "sections.0",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/1563036",
                "endSection": "sections.0",
                "offsetInBeginSection": 0,
                "offsetInEndSection": 162,
                "text": "Disease patterns in RA vary between the
                sexes; the condition is more commonly seen in
                women, who exhibit a more aggressive disease
                and a poorer long-term outcome."
            },
            {
                "beginSection": "sections.2",
                "document": "http://www.ncbi.nlm.nih.gov/pubmed/12723987",
                "endSection": "sections.2",
                "offsetInBeginSection": 7,
                "offsetInEndSection": 89,
                "text": "The mean age of the patients was 62 years
                (range 19\u201396 years) and 71% were female;"
            },
            {
```

```
        "beginSection": "sections.0",
        "document": "http://www.ncbi.nlm.nih.gov/pubmed/15083883",
        "endSection": "sections.0",
        "offsetInBeginSection": 688,
        "offsetInEndSection": 830,
        "text": "The female to male ratio was 2.5:1 and the mean
        age at diagnosis was 49.4 +/- 14.9 years for women and
        55.3 +/-15.6 years for men (P < 0.0003)"
    },
    {
        "beginSection": "sections.0",
        "document": "http://www.ncbi.nlm.nih.gov/pubmed/16418123",
        "endSection": "sections.0",
        "offsetInBeginSection": 453,
        "offsetInEndSection": 514,
        "text": "in 244 female and 91 male patients with rheumatoid
        arthritis."
    }
]  "type": "factoid"
}
```

A.3 Fragment of BioASQ dataset

In our testing data, this question has no related snipptes.

```
"questions": [
        {
"body": "Does metformin interfere thyroxine absorption?",
         "concepts": [
"http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?
 field=uid&exact=Find+Exact+Term&term=D008687",
 "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?
  field=uid&exact=Find+Exact+Term&term=D013974",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?
  field=uid&exact=Find+Exact+Term&term=D000042"
],
"documents": [
"http://www.ncbi.nlm.nih.gov/pubmed/23554450",
"http://www.ncbi.nlm.nih.gov/pubmed/23264396",
"http://www.ncbi.nlm.nih.gov/pubmed/23244059",
"http://www.ncbi.nlm.nih.gov/pubmed/23154888",
"http://www.ncbi.nlm.nih.gov/pubmed/23072197",
"http://www.ncbi.nlm.nih.gov/pubmed/21748540",
"http://www.ncbi.nlm.nih.gov/pubmed/21633823",
"http://www.ncbi.nlm.nih.gov/pubmed/21435090",
"http://www.ncbi.nlm.nih.gov/pubmed/21468525",
"http://www.ncbi.nlm.nih.gov/pubmed/21041167"
],
"exact_answer": "No",
"id": "51406e6223fec90375000009",
"ideal_answer": [
"There are not reported data indicating that
 metformin interferes with thyroxine absorption"
],
"type": "yesno"
}
```

References

Abney, S. P. (1991). Parsing by chunks. In *Principle-Based Parsing* (pp. 257–278). Kluwer Academic Publishers, Dordrecht.

Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, *33*, 157–177.

Alguliev, R. M., & Aliguliyev, R. M. (2005). Effective summarization method of text documents. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on* (pp. 264–271). Institute of Electrical and Electronics Engineers.

Ananiadou, S., Kell, D. B., & Tsujii, J.-i. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, *24*, 571–579.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium* (p. 17). American Medical Informatics Association.

Battiti, R., & Brunato, M. (2014). *The LION Way: Machine Learning plus Intelligent Optimization*. Lionsolver Inc., Los Angeles, June 2013.

Becker, K. G., Hosack, D. A., Dennis, G., Lempicki, R. A., Bright, T. J., Cheadle, C., & Engel, J. (2003). Pubmatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, *4*, 61.

Berend, G. (2011). Opinion expression mining by exploiting keyphrase extraction. In *International Joint Conference on Natural Language Processing* (pp. 1162–1170).

BioASQ (2014a). The bioasq challenges. `http://www.bioasq.org`. Accessed Oct 9, 2014.

BioASQ (2014b). Question answering from texts and structured information. `http://www.bioasq.org/participate/how_bioasq_relates_to_your_research#4`. Accessed Oct 8, 2014.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., & Salakoski, T. (2009). Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* (pp. 10–18). Association for Computational Linguistics.

Blaschke, C., Andrade, M. A., Ouzounis, C. A., & Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. In *Intelligent Systems for Molecular Biology* (pp. 60–67). volume 7.

Bui, Q.-C., & Sloot, P. M. (2012). A robust approach to extract biomedical events from literature. *Bioinformatics*, *28*, 2654–2661.

Chowdary, C. R., Sravanthi, M., & Kumar, P. S. (2010). A system for query specific coherent text multi-document summarization. *International Journal on Artificial Intelligence Tools*, *19*, 597–626.

Cohen, A. (2004). Using symbolic network logical analysis as a knowledge extraction method on medline abstracts. In *Oregon Health Sciences University Scholar Archive*. Oregon Health Sciences University.

Cohen, A. M. (2005). Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the Acl-ismb Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics* (pp. 17–24). Association for Computational Linguistics.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. (pp. 57–71). Oxford University Press volume 6.

Dostál, M., & Jezek, K. (2011). Automatic keyphrase extraction based on nlp and statistical methods. In *Databases, Texts, Specifications, and Objects* (pp. 140–145).

EBM-Group (1992). Evidence-based medicine. a new approach to teaching the practice of medicine. *JAMA: the Journal of the American Medical Association*, *268*, 2420.

Eddy, D. M. (2005). Evidence-based medicine: a unified approach. *Health Affairs*, *24*, 9–17.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the Association for Computing Machinery (JACM)*, *16*, 264–285.

Frank, E., Paynter, G. W., Witten, I. H., & Gutwin, C. (1999). Domain-specific keyphrase extraction. In *Sixteenth International Joint Conference on Artificial Intelligence*.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, *3*, 115–130.

Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P., & Cöster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*, *67*, 49–61.

Friedman, C., & Liu, H. (2002). Mining terminological knowledge in large biomedical corpora. In *Pacific Symposium on Biocomputing 2003: Kauai, Hawaii, 3-7 January 2003* (p. 415). World Scientific.

Giarlo, M. J. (2005). A comparative analysis of keyword extraction techniques. CiteSeer.

Glasziou, P., Del Mar, C., & Salisbury, J. (2010). Evidence based medicine workbook. British Medical Journa Group.

Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, *2*, 258–268.

Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., & Frank, E. (1999). Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, *27*, 81–104.

Hammouda, K. M., Matute, D. N., & Kamel, M. S. (2005). Corephrase: Keyphrase extraction for document clustering. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 265–274). Springer.

Hersh, W. R., Crabtree, M. K., Hickam, D. H., Sacherek, L., Friedman, C. P., Tidmarsh, P., Mosbaek, C., & Kraemer, D. (2002). Factors associated with success in searching medline and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association*, *9*, 283–293.

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, *6*.

Holzinger, A., Geierhofer, R., & Errath, M. (2007). Semantische informationsextraktion in medizinischen informationssystemen. *Informatik-Spektrum*, *30*, 69–78.

Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). Biomedical text mining: State-of-the-art, open problems and future challenges. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics* (pp. 271–300). Springer.

Hovy, E., & Lin, C.-Y. (1998). Automated text summarization and the summarist system. In *Proceedings of a Workshop on held at Baltimore, Maryland: October 13-15, 1998* (pp. 197–214). Association for Computational Linguistics.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 216–223). Association for Computational Linguistics.

Hulth, A., & Megyesi, B. B. (2006). A study on automatically extracted keywords in text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 537–544). Association for Computational Linguistics.

Hussey, R., Williams, S., & Mitchell, R. (2012). Automatic keyphrase extraction: a comparison of methods. (pp. 18–23). eKNOW 2012, The Fourth International Conference on Information, Process, and Knowledge Management.

Jin, F., Huang, M., & Zhu, X. (2009). A query-specific opinion summarization system. (pp. 428–433). Institute of Electrical and Electronics Engineers Cognitive Informatics, 2009. ICCI'09. 8th IEEE International Conference on.

Jones, K. S. et al. (1999). Automatic summarizing: factors and directions. (pp. 1–12). Cambridge, MA: MIT Press.

Kaur, J., & Gupta, V. (2010). Effective approaches for extraction of keywords. (p. 6). volume 7.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., & Tsujii, J. (2009). Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* (pp. 1–9). Association for Computational Linguistics.

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*, i180–i182.

Krulwich, B., & Burkey, C. (1996). Learning user information interests through extraction of semantically significant phrases. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access* (pp. 100–112).

Larkey, L., AbdulJaleel, N., & Connell, M. (2003). Whats in a name?: Proper names in arabic cross language information retrieval. In *ACL Workshop on Computing Approaches to Semitic Languages*. Citeseer.

Lee, S., & Kim, H. (2008). News keyword extraction for topic tracking. In *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on* (pp. 554–559). IEEE volume 2.

Lingeman, J., & Dietz, L. (2014). Umass at bioasq 2014: Figure-inspired text retrieval. In *2nd BioASQ Workshop: A challenge on large-scale biomedical semantic indexing and question answering*. BioASQ QA Track 2014 Workshop Programme.

Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization* (pp. 17–24). Association for Computational Linguistics.

Liu, F., Liu, F., & Liu, Y. (2008). Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In *Spoken Language Technology Workshop, 2008. spoken-language technologies 2008. IEEE* (pp. 181–184). IEEE.

Liu, F., Pennell, D., Liu, F., & Liu, Y. (2009). Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 620–628). Association for Computational Linguistics.

Liu, H., Christiansen, T., Baumgartner Jr, W. A., & Verspoor, K. (2012). Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *JournalBiomedical Semantics*, *3*, 17.

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*, *37*, 1–41.

Lott, B. (2012). Survey of keyword extraction techniques. The University of New Mexico Education.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*, 159–165.

Ma, L., He, T., Li, F., Gui, Z., & Chen, J. (2008). Query-focused multi-document summarization using keyword extraction. In *Computer Science and Software Engineering, 2008 International Conference on* (pp. 20–23). IEEE volume 1.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, *13*, 157–169.

McKibbon, K. (1998). Evidence-based practice. *Bulletin of the Medical Library Association*, *86*, 396.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into texts. Association for Computational Linguistics.

Miwa, M., Thompson, P., & Ananiadou, S. (2012). Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, *28*, 1759–1765.

Paice, C. D. (1980). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval* (pp. 172–191). London: Butterworths.

Paliouras, G., & Krithara, A. (2015). Bioasq final report. (pp. 7–18). BioASQ Group.

Pamela Corley, E. E., & Adrian Follette, E. W. (2003). ksom. `http://www.usc.edu/hsc/nml/portals/orientation/`. Accessed Apr 23 , 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

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

Pustejovsky, J., Castano, J., Sauri, R., Rumshinsky, A., Zhang, J., & Luo, W. (2002). Medstract: creating large-scale information servers for biomedical libraries. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical domain-Volume 3* (pp. 85–92). Association for Computational Linguistics.

Qazvinian, V., Radev, D. R., & Özgür, A. (2010). Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 895–903). Association for Computational Linguistics.

Qiu, M., Li, Y., & Jiang, J. (2012). Query-oriented keyphrase extraction. In *Information Retrieval Technology* (pp. 64–75). Springer Berlin Heidelberg.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, *28*, 399–408.

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural Language Processing using very large corpora* (pp. 157–176). Springer.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, *60*, 503–520.

Sackett, D. L., Rosenberg, W. M., Gray, J., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ: British Medical Journal*, *312*, 71.

Sarkar, K., Nasipuri, M., & Ghose, S. (2010). A new approach to keyphrase extraction using neural networks. *International Journal of Computer Science Issues Publicity Board 2010*, .

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21.

Suematsu, H. (1994). Current status of the edr electronic dictionary project and its evaluation research. *New Generation Computing*, *12*, 311–315.

Tanabe, L., & Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics*, *18*, 1124–1132.

Tanabe, L., Xie, N., Thom, L. H., Matten, W., & Wilbur, W. J. (2005). Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, *6*, S3.

Ulrich, R. S., Zimring, C., Zhu, X., DuBose, J., Seo, H.-B., Choi, Y.-S., Quan, X., & Joseph, A. (2008). A review of the research literature on evidence-based healthcare design. *HERD: Health Environments Research & Design Journal*, *1*, 61–125.

Veenstra, J., & Buchholz, S. (1998). Fast np chunking using memory-based learning techniques. In *Proceedings of BENELEARN98* (pp. 71–78). Citeseer.

Vlachos, A., & Craven, M. (2012). Biomedical event extraction from abstracts and full papers using search-based structured prediction. *BMC Bioinformatics*, *13*, 1–11.

Wan, X., & Xiao, J. (2008a). Collabrank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 969–976). Association for Computational Linguistics.

Wan, X., & Xiao, J. (2008b). Single document keyphrase extraction using neighborhood knowledge. In *Association for the Advancement of Artificial Intelligence* (pp. 855–860). volume 8.

Wang, J., Peng, H., & Hu, J. (2006). Automatic keyphrases extraction from document using neural network. In *Advances in Machine Learning and Cybernetics* (pp. 633–641). Springer.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). Kea: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries* (pp. 254–255). Association for Computing Machinery.

Yu, H., Hripcsak, G., & Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*, *9*, 262–272.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, *4*, 1169–1180.