Automated Medical Text Summarisation to Support Evidence-based Medicine

ABEED SARKER

This thesis is presented for the degree of
Doctor of Philosophy
at the
Centre for Language Technology
Department of Computing
Macquarie University
NSW 2109 Australia

Submitted December 2013
Completed June 2014

Abstract

Clinical guidelines urge medical practitioners to perform Evidence-based Medicine: a practice that requires practitioners to incorporate the best available evidence from published research and from clinical practice, when making decisions. Due to the abundance of published medical research available, practitioners often fail to follow evidence-based guidelines at point-of-care due to time constraints. As such, this practice can vastly benefit from automatic systems that can generate short, reliable, evidence-based summaries in response to queries posed by the practitioners. Analysis of a corpus specialised for text summarisation in the evidence-based medicine domain suggests that a summarisation system tailored to this domain must be capable of performing two tasks: (i) query-focused text summarisation and (ii) automatic appraisal of the evidence. In this thesis, we utilise data from a specialised corpus to address these two facets of the problem. Our investigations lead to the following three contributions/findings:

1. A *model for the automatic generation of evidence grades*: We show that a supervised classification model can be used to perform automatic quality grading of evidence on a chosen scale. Rule-based approaches can be applied to the text of the medical articles to extract useful features, which can be utilised in the classification task. Using a sequence of high precision machine learning classifiers, we achieve recall and grading accuracies that are comparable to human performance, and significantly better than baseline systems.

2. An *approach for the summarisation of information from individual medical texts*: We use a simple extractive summarisation model, which attempts to identify salient sentences by scoring them based on various statistics. We show that summarisation performance can be significantly improved by deriving statistics from the specialised corpus, applying *target-sentence-specific* scoring, incorporating query-associated information, and incorporating domain knowledge in several ways. Our extractive summarisation system outperforms multiple baseline and benchmark systems in our automatic evaluations with a percentile rank of 96.8%.

3. A possible *approach for the generation of bottom-line evidence-based summaries using individual single-document summaries*: Based on our investigations of human authored summaries and our automatic single-document summaries, we show that content-rich

single-document summaries may be used for the generation of bottom-line recommendations from multiple documents. We apply sentence level polarity classification on the single-document summary sentences and show that automatically identified context-sensitive polarities may be used to make bottom-line recommendations regarding interventions.

The thesis describes our corpus-based investigation of the evidence-based summarisation process. Our development and evaluation of models for the core facets of evidence-based summarisation fill some of the gaps in end-to-end medical question answering research. In executing this research, we contribute to our general understanding of medical text summarisation approaches.

Statement of Candidate

I certify that the work in this thesis entitled **"Automated Medical Text Summarisation to Support Evidence-based Medicine"** has not been previously submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis.

Abeed Sarker
December 4, 2013

Acknowledgements

Firstly, I thank my supervisors Diego and Cecile for their support throughout this degree and for providing me with much needed guidance. I would also like to express my gratitude to the other academics of the Department of Computing and my colleagues at the Centre for Language Technology.

I would also like to take this opportunity to thank my parents and my brother. I have been lucky enough to have two pairs of parents: I thank the first pair for my birth and my wonderful childhood and the second for dealing with my childlike behaviour as an adult. I particularly thank my birth parents for always making me feel special and for believing in me.

I am indebted to my friends: they are no less important to me than my family. Without them I would not be able to make it through my Bachelors, Honours and PhD degrees.

I am indebted to all the musicians who have entertained me throughout my life and the duration of this degree. I would like to take this opportunity to thank all of them for keeping me alive. I would also like to thank the current and past players of my favourite football club A. C. Milan for entertaining me throughout the umpteen years of my 'fanhood'.

I would like to convey a special acknowledgement to all the activists and freedom fighters of Bangladesh who selflessly sacrificed their lives for independence in 1971. Without their sacrifice, I would not have an identity, let alone a PhD degree.

Finally, I would like to thank this wonderful country (Australia). I have spent my entire adult life here, and in it I have found a home away from home.

Publications

This thesis is based on research I have performed with the help of my supervisors and other colleagues during my PhD degree at the Department of Computing, Macquarie University between 2010 and 2013. Some parts of the thesis include revised versions of the following papers published as a result of the doctoral work undertaken by me:

- A. Sarker, D. Mollá-Aliod, and C. Paris, "Automatic Prediction of Evidence-based Recommendations via Sentence-level Polarity Classification," *in Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, (Nagoya, Japan), 2013.
- A. Sarker, D. Mollá, and C. Paris, "An Approach for Query-focused Text Summarisation for Evidence Based Medicine," *in Proceedings of the 14th conference on Artificial Intelligence in Medicine (AIME)*, (Murcia, Spain), 2013.
- A. Sarker, D. Mollá-Aliod, and C. Paris, "An Approach for Automatic Multi-label Classification of Medical Sentences," *in Proceedings of LOUHI 2013: Fourth International Workshop on Health Document Text Mining and Information Analysis*, (Sydney, Australia), 2013.
- A. Sarker, D. Mollá-Aliod, and C. Paris, "Towards Two-step Multi-document Summarisation for Evidence Based Medicine: A Quantitative Analysis," *in Proceedings of the Australasian Language Technology Association (ALTA) Workshop*, (Dunedin, New Zealand), 2012.
- A. Sarker, D. Mollá and C. Paris, "Extractive summarisation of medical documents using domain knowledge and corpus statistics," *Australasian Medical Journal 2012, 5, 9*, 2012.
- A. Sarker, D. Mollá and C. Paris, "Extractive Evidence Based Medicine Summarisation Based on Sentence-Specific Statistics," *in Proceedings of the 25th International Symposium on Computer-Based Medical Systems*, (Rome, Italy), 2012.
- A. Sarker, D. Mollá-Aliod and C. Paris, "Extractive Summarisation of Medical Documents using Domain Knowledge and Corpus Statistics," *in Proceedings of the First Australian Workshop on Artificial Intelligence in Health*, (Perth, Western Australia, Australia), 2011.

- D. Mollá-Aliod and A. Sarker, “Automatic Grading of Evidence: the 2011 ALTA Shared Task," *in Proceedings of the Australasian Language Technology Association (ALTA) Workshop*, (Canberra, Australian Capital Territory, Australia), 2011.
- A. Sarker, D. Mollá-Aliod, and C. Paris, “Towards automatic grading of evidence,” *in Proceedings of LOUHI 2011: Third International Workshop on Health Document Text Mining and Information Analysis*, (Bled, Slovenia), 2011.
- A. Sarker and D. Mollá-Aliod, “A Rule-based Approach for Automatic Identification of Publication Types of Medical Papers," *in Proceedings of the Fifteenth Australasian Document Computing Symposium*, (Melbourne, Victoria, Australia), 2010.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

Evidence-based medicine is a practice that requires healthcare practitioners to obtain the best quality clinical evidence from published medical research when answering clinical queries, in addition to using their own clinical expertise [Sackett et al., 1996]. Due to the plethora of available online text-based medical research articles, healthcare practitioners usually face the problem of information overload when attempting to search for, appraise, and extract information from relevant medical literature. Research has shown that practitioners often fail to pursue evidence-based answers to their clinical queries primarily due to time related constraints at point-of-care [Ely et al., 1999, 2005]. As such, there is a strong motivation for text processing systems that can automate some of the processes involved in this practice. However, unlike some other domains (e.g., the news domain [Sarker et al., 2013]), research on automatic text processing in the medical domain is still very much in its infancy. The intent of our research is to identify automatic approaches for performing some tasks associated with evidence-based medicine practice. In particular, we explore the possible automation of the processes involved in the generation of final recommendations in response to clinical queries, given the set of relevant source documents associated with a query.

Our analysis of a corpus, containing a large number of medical questions and manually authored evidence-based answers to these questions, reveals the various important aspects of an evidence-based answer, and also the manual answer generation process. The generation of an evidence-based answer is a relatively complicated process and begins when a healthcare practitioner is faced with a clinical question during everyday practice. The practitioner is required to formulate a *query* which represents his or her information need, and search a database to obtain relevant research articles. Once the articles are retrieved, the practitioner must *extract* the important information relating to the question posed and *synthesise* the information from different publications on the

topic to produce a short, specific recommendation. Simultaneously, the practitioner must appraise the extracted information to *identify the quality* of the evidence behind the given answer.

Figure 1.1 provides an example of the result of the step-by-step evidence-based answer generation approach. The figure shows the question posed by the practitioner, the answers extracted from various research articles based on the question, a bottom-line answer produced by manually synthesising the different answers utilising domain expertise, and a grade indicating the quality of the evidence from which the answer was generated. The query posed attempts to identify the *best strategy for impaired glucose tolerance in non-pregnant adults*. Two answers are shown in the example which were obtained from distinct research papers (the PubMed ID indicates their identification number in PubMed[1]). Both papers indicate that lifestyle interventions work best for this problem. These two answers are manually combined to produce the synthesised answer which recommends several lifestyle interventions. Finally, the quality of the evidence associated with the answer is presented, which in this case is B[2] on a chosen scale [Ebell et al., 2004]. This data comes from the 'Clinical Inquiries' section of the Journal of Family Practice[3], which is a popular source of real life data from evidence-based medicine practice.

Based on our observation of the data collected from evidence-based medicine practice, we formulate the process of generating evidence-based answers as a two-step summarisation process accompanied by a simultaneous sub-process that assesses the quality of the grade of the evidence. When formulated in terms of input and output, the first step of the summarisation process can be defined as follows:

Input-1: A clinical query
Input-2: A source article
Output: A summary of *Input-2* based on the information needs of *Input-1*

Similarly, the second step can be formulated as a summarisation process with inputs and the output as follows:

Input-1: A clinical query
Input-2: Multiple summarised source articles
Output: A multi-document summary of *Input-2* based on the information needs of *Input-1*

[1] http://www.ncbi.nlm.nih.gov/pubmed/. Accessed on 26th May, 2014.
[2] The scale and algorithms for determining the quality of an evidence are described later in the thesis.
[3] www.jfponline.com/. Accessed on 26th May, 2014.

Question.
What is the best strategy for impaired glucose tolerance in nonpregnant adults?

Extracted Answer 1.
In the Finnish Diabetes Prevention Study, the lifestyle intervention group had a 12.5% absolute risk reduction compared with the control group (NNT=8). Successful lifestyle interventions included weight loss of 5% to 7%, decreased fat intake, increased fiber intake, and 150 minutes of exercise per week. [*PubMed ID: 11333990*]

Extracted Answer 2.
In the US Diabetes Prevention Program (DPP), 3234 patients with IGT and a body-mass index (BMI) of at least 24 kg/m^2 were randomly assigned to one of the following groups: placebo, metformin, or intensive lifestyle modification. After an average follow-up of 2.8 years, there was a 14% absolute risk reduction in the progression to diabetes in the lifestyle intervention group compared with placebo (number needed to treat [NNT]=7). Successful lifestyle interventions included weight loss of 5% to 7%, decreased fat intake, increased fiber intake, and 150 minutes of exercise per week. [*PubMed ID: 11832527*]

Synthesised Answer.
The best treatment strategy for impaired glucose tolerance IGT and impaired fasting glucose IFG is lifestyle intervention with a structured weight loss program of diet and exercise. Patients with IGT and IFG should be counseled to lose 5% to 7% of their body weight and instructed on moderate intensity physical activity for 150 minutes per week.

Grade for Quality of Evidence: B

Figure 1.1: The various components of an evidence-based answer: A clinical question taken from evidence-based medicine practice, two brief answers obtained from research articles to answer the clinical question, a synthesised bottom-line answer, and a grade for the quality of the evidence.

Finally, the quality of the evidence determination process can also be formulated in terms of input and output as follows:

Input-1: A clinical query
Input-2: The source articles associated with an evidence-based answer to the query
Output: A grade on some chosen scale indicating the quality of the evidence obtained for *Input-1* from *Input-2*

Thus, the whole process involves information extraction from text, summarisation of information to generate bottom-line recommendations, and analysis of lexical information to deduce the qualities of the extracted evidences. While the first two steps of summarisation need to be completed in order, the process of generating a grade of evidence for the overall, multi-document summary can be carried out independently of the other two processes. This is because the grade of the evidence from which an answer is derived depends largely on the qualities of the publications containing the evidence. In our model (explained in detail in Chapter 3), therefore, we attempt to predict the quality grade of the final evidence from the source texts associated with it.

In this research work, we address the problem of automatic summarisation of medical text for evidence-based medicine. In particular, we address the three related but distinct problems: *information extraction*, *generation of bottom-line answers*, and *grading of the quality of evidence*.

1.2 Text Summarisation and Evidence-based Medicine

In this section we introduce some of the concepts of Natural Language Processing (NLP) that are associated with evidence-based medicine. More specifically, we explain how the problem of generating *evidence-based summaries* is essentially a text summarisation problem specialised to this domain. The discussion in this section is intentionally kept brief, and the literature associated with some of the discussion we present here is cited and elaborated in the next chapter.

1.2.1 Single and Multi-document Text Summarisation

Automatic text summarisation approaches can be classified based on the number of input documents/items [Mani, 2001]. Single-document summarisation is the simplest case, where the purpose of the summary is to present the important information in the document in a condensed manner. The problem, therefore, is to identify important content in a document and present it as the summary. The problem of multi-document summarisation is more complex than simply finding and presenting text segments. When summarising information from multiple documents, the summariser has to make decisions about the ordering and synthesis of information from multiple documents. For the summaries to be coherent, information must be presented in the correct order; for the summaries to be precise and concise, presenting redundant information from different documents must be avoided; finally, inconsistent information present in different documents must be identified and dealt with. Understandably, summarising multiple documents automatically is a much more challenging task and requires more customisation based on the domain.

1.2.2 Extractive and Abstractive Summarisation

Extractive summaries are simply text extracts that are presented as summaries. Text extracts can be words, phrases, sentences or text windows of predefined sizes. While such approaches are sufficient for extracting and presenting relevant content, the extracts are generally a list of disjoint textual units [Sparck Jones, 1999, 2007]. As such, they may lack sufficient information, be grammatically incorrect and lead to possible misinterpretation. Alternative methods generate summaries that have proper structure and are usually grammatically correct. Such summaries may be generated by organising extracted material into a deliberate form, or by generating novel text from conceptual or other forms of representation (e.g., a lexical graph). Such summaries are called *abstractive* summaries, and this is the type of summaries that humans generally produce. The automatic generation of abstractive summaries generally requires additional processing compared to the generation of extractive summaries.

1.2.3 Incorporating Query-focus in Summarisation

The content of an automatically generated summary can be influenced by the presence of a query [Mani, 2001, Sparck Jones, 2007]. Unlike generic summarisation, where the summary presents the key information from one or more documents, when summarising based on a query, only information relevant to the query is extracted and summarised. Thus, the summarisation approaches that take queries into account are more targeted towards satisfying the information need of the query. Such approaches are called *query-focused* summarisation approaches.

1.2.4 Summarisation and Text Classification

Text classification is the task of assigning bodies of text (e.g., documents) to distinct *classes* or *categories* based on the content of the text [McCallum and Nigam, 1998, Kilicoglu et al., 2009, Kim et al., 2011]. Content based classification techniques generally attempt to extract relevant content from the text bodies (e.g., words, phrases and so on) and use that information as *features* for deciding the category or class for that body of text. Text classification approaches have been applied to solve various NLP problems including summarisation, where text segments are classified as summary sentences or non-summary sentences. Text classification has also been applied, in various domains including medical, for categorising text in terms of quality of information.

The tasks described in this thesis attempt to let the information seeker (*i.e.,* the healthcare practitioner) specify the information needs through a query. Based on the information needs expressed, our approaches attempt to perform various tasks associated with evidence-based summarisation,

including single- and multi-document summarisation, and the automatic prediction of evidence grades.

1.2.5 Formalising Evidence-based Summarisation

We have suggested that the preparation of evidence-based answers can be thought of as a step-by-step summarisation process which starts when the information needs are expressed by the practitioner in the form of a query. The first step of identifying and extracting content from individual medical documents can be modelled as a query-focused, single-document summarisation process, which may be abstractive or extractive based on the needs. Since, in our case, the purpose of this step is to extract key information from source documents, an extractive summarisation model appears appropriate. The task for this step is to identify the key information and extract it. The second step, which involves synthesising information from multiple documents based on the information needs of the query, can be modelled as query-focused, multi-document summarisation. Again, this can be extractive or abstractive, but is likely to be the latter. This is because the intent of this step is to generate concluding recommendations based on the information provided, which requires the synthesis information from each of the individual extracts. Finally, the task of appraisal of medical articles relevant to a specific query to estimate a grade for the quality of evidence can be modelled as a text classification problem. The grade is not assigned to single documents, instead it is based on a number of features which are obtained from all the relevant input documents combined (i.e., the full set of evidence). Since this process relies on the source documents from which the evidence is obtained, it is not necessarily reliant on the summarisation process and can be seen as a parallel process.

Thus, our formulation of the process of generating evidence-based answers to clinical queries involves automatic text summarisation at different granularities and automatic text classification. Combining the inputs and outputs associated with the tasks, we model our approaches based on the following assumptions:

1. The task of evidence-based answer generation involves information extraction, appraisal and synthesis.

2. The tasks of grading the quality of evidence and summarising information take place in parallel to each other and the techniques employed for each of the tasks complement each other.

Importantly, in real life evidence-based medicine practice, the outputs for each of these processes are reliant solely on the text-based inputs, but, in practice, the transformation of input texts into evidence-based summaries are performed using the domain expertise of the practitioners. Thus,

automating these techniques is challenging since vast amounts of domain knowledge must be incorporated. At the same time, automation is desirable due to the amount of valuable time it can save. We provide some estimates about the time required to manually perform some of the tasks associated with evidence-based medicine practice later in the thesis, which indicate how much time can be saved by the automation of these tasks.

1.3 Research Questions Addressed

A computational account of evidence-based summary generation requires techniques for the two core issues: generating summary content and generating quality grades for evidence. As mentioned in the previous section, generating summary content can be viewed as a two-step summarisation process. In the first step, relevant content is summarised from individual source documents based on the focus of a given query. This single-document summarisation stage specifies the importance of various text segments that can be utilised by another multi-document summary generation step to provide short, bottom-line answers. Assessing the quality of evidence is a simultaneous process, and the qualities are determined by the practitioners during the answer generation process. We, therefore, focus on three related problems in this thesis: generating grades for the qualities of evidence, summarising individual documents based on the information needs of a query, and generating bottom-line answers to clinical queries as a multi-document summarisation process.

Solutions to these three problems affect the eventual output of the summarisation system: a short, evidence-based summary and a grade for the quality of the evidence of the summary. However, simply evaluating the end product of the summarisation system obscures our understanding of how solutions to the three problems affect the final summary. Accordingly, to study each problem effectively, we investigate the three components of the summarisation system — these are described in more detail in Chapters 4, 5, and 6. We evaluate these components in isolation from each other, using appropriate data sets, to minimise the number of confounding factors.

As a consequence of this methodology, we structure the thesis to present answers to the core research questions in evidence-based summarisation. We discuss the general problem of evidence-based summarisation and its functionality in an end-to-end clinical question answering system, but do not attempt to create or evaluate such a system. It would require solutions to additional but peripheral problems such as information retrieval — which is the research area that focuses on obtaining information resources relevant to information needs from collections of information sources. However, exploration of such peripheral research areas is outside the scope of the research described in this thesis. Instead, we focus on the three problems which are key to evidence-based summarisation; we describe these now in more detail.

1.3.1 How can the various properties of medical text be utilised to automatically appraise the quality of evidence on a specialised scale?

The grade for the quality of evidence accompanying a query-focused summary depends on a number of factors that involve each of the individual source texts associated with the query, and also all the texts collectively. In our corpus, the possible grades that a summary can have are **A**, **B**, and **C**, which indicate *high*, *moderate*, and *low* quality answers respectively, according to the Strength Of Recommendation Taxonomy (SORT) [Ebell et al., 2004]. Using this as our target scale and the gradings provided in our corpus as target grades, we first explore and identify important features of the source texts that influence the grades of evidence. The features we analyse include those that are embedded within the texts and also meta-data associated with the source texts. We first examine the extent to which various features influence evidence grades. We apply rule-based approaches to automatically extract some of the features from the source texts, and then model the problem of evidence-grading as a text classification problem using our extracted features. We measure the performance of our approach by comparing its accuracy with some baseline systems, and also by comparing it with human agreement for this problem.

1.3.2 What approaches can be used to generate content-rich, query-focused summaries from single documents?

When provided with a query and a source text, the first problem is to select informative content from the text based on the information needs of the query. The selected content must be short and yet have sufficient coverage of the relevant information. In this thesis, we explore extractive summarisation approaches for the task of content selection. In our approach, we consider sentences as extractable text segments and produce three sentence extractive summaries from the source texts. Our summarisation approach is data-driven, since we utilise various statistics from a specialised corpus. We develop a target-sentence-specific ranking mechanism to perform extractive summarisation, combine a range of surface, intra-sentence and inter-sentence features, and incorporate medical domain knowledge in various ways to assign scores to sentences. Our approach produces a ranked list of sentences for each document. Also, since the lengths of the target summaries are known (in our case, we use summary lengths of three sentences in line with related research in this domain [Lin and Demner-Fushman, 2007]), we apply separate scoring mechanisms for each target sentence (i.e., first, second, or third). This allows us to focus on different types of lexical content with each of the summary sentences (i.e., each of the three summary sentences have a different purpose). We evaluate our approach relative to existing baselines and benchmark approaches, and show that our extractive approach is superior.

1.3.3 How can a bottom-line recommendation be generated by combining information from multiple documents?

Given a clinical query, there are generally multiple documents providing relevant information that can be used to answer the query. To generate a bottom-line recommendation, the relevant information must be synthesised to assess the information presented in them collectively. The synthesis of information is perhaps the most difficult because it requires practitioners to utilise their domain expertise and judge the contribution of each individual evidence. The final recommendations have to take into account various information such as the effectiveness of interventions for disorders, side-effects, types of interventions, etiologies, prognoses, management procedures, general suggestions and so on. Due to the complex nature of this task, it is possible that experts often mostly rely on their own expertise for the final decision making, rather than extracting evidence from literature. Therefore, we commence our work in this area by first performing an automatic analysis to estimate the extent to which experts use information from the literature to generate the bottom-line recommendations. We do this by defining several techniques to measure the degree to which information in the bottom-line summaries in our corpus are *covered* by information contained in the source documents. We also apply the same coverage measurements to determine the validity of a two-step summarisation model and the possible advantages and disadvantages of such an approach. We analyse the applicability of two popular existing multi-document summarisation techniques to generate bottom-line recommendations, and propose an approach using automatic, context-sensitive sentence level polarity classification. Finally, we discuss possible future directions for the generation of bottom-line summaries using sentence level polarity classification approaches.

1.4 Research Findings

We focus on the three questions above in order to provide an account of evidence-based answer generation as a specialised text summarisation process. Our investigations lead to a number of research findings that constitute novel contributions to the field of medical text summarisation, and, therefore, to the broader field of computational linguistics. These include novel content selection and ranking approaches that have potential applications to other domains, an approach for grading the quality of evidence via text classification, and approaches for performing multi-document summarisation in this domain. We now briefly present descriptions of the three contributions and a summary of corresponding evaluation results. These are described in detail in subsequent chapters.

1.4.1 An Approach for Automatic Evidence Grading

Using supervised machine learning, we identify important features that influence evidence-grades. Applying rule-based approaches that use regular expressions, we automatically extract certain important features from the source texts, such as the *Publication Types* of the articles. Our experiments show that such an approach is effective. We combine computationally extracted features, and features collected directly from the meta-data associated with articles to automatically estimate the grades of evidence-based summaries. We propose a classification strategy that utilises a sequence of high precision classifiers to perform the classification task. This leads to our first research finding:

Research Finding 1. *Grades indicating the qualities of clinical evidence can be automatically predicted through the use of supervised machine learning and informative features. The structure and content of medical abstracts can be exploited by rule-based systems to extract important features to be used in the supervised machine learning process. We show that by applying carefully selected features, supervised machine learning can be used for evidence grading. We also show that our grading system outperforms our proposed baseline and performs comparably to humans on the same data set.*

1.4.2 Extractive Summarisation for Content Selection and Ranking

We apply a classic sentence scoring approach which uses a linear equation of weighted feature scores for scoring each sentence. We introduce several novelties such as the use of target-sentence-specific scoring (i.e., applying separate scores depending on the position of the target sentence in the summary). In addition to using various statistics from our specialised corpus, we also incorporate domain knowledge in a number of ways including classification techniques for identifying the rhetorical status of sentences. Furthermore, we customise the summarisation approach to the *type* of question posed and show that considering this information improves the qualities of extractive summaries. This leads to our second finding:

Research Finding 2. *The qualities of extractive summaries in this domain can be improved by utilising statistics from a specialised corpus. The availability of data from a specialised corpus enables the derivation of various statistics from seen data, rather than intuitions about text in the medical domain. We also customise the scoring approach to the types of questions and show that incorporating that information in the summary generation process can benefit summarisation performance. Additionally, we apply a target-sentence-specific scoring strategy, as opposed to a generic scoring strategy which gives the same feature score irrespective of the position of the target sentence in the summary. A target-sentence-specific scoring strategy overcomes the problem of underfitting that approaches using a generic scoring strategy face. Our extractive summarisation approach outperforms several well-established baselines within our domain and*

also several domain-independent summarisation techniques when evaluated on a relative scale.

1.4.3 Multi-document Summarisation for Answer Generation

We perform an in-depth analysis on a corpus specialised in text summarisation for evidence-based medicine and conclude that it may be useful to consider the summary generation process to be a two-step process, where the second step involves synthesising information from the output of the query-focused, single-document summaries mentioned above. We define some quantitative measures to estimate the validity of such an approach. We also assess the performance of two existing abstractive, multi-document summarisation approaches to this task of information synthesis and show why some approaches are likely to work while others are not. Finally, we show that automatic, sentence level polarity classification techniques can be applied to generate recommendations relative to interventions, which in turn can be used to generate bottom-line recommendations. These analyses lead to our final research finding:

Research Finding 3. *A significant proportion of the contents of the bottom-line recommendations prepared by evidence-based medicine practitioners comes from the available medical literature. Thus, the generation of bottom-line recommendations in response to a clinical query may be performed automatically via text summarisation approaches. Such an answer generation technique may benefit from the use of a two-step process, if the first step is capable of extracting content-rich text and discarding noise. A possible way to use the single-document extractive summaries is to apply context-sensitive polarity detection techniques.*

1.5 Thesis Outline

Chapter 2 provides a detailed overview of relevant literature. The review is divided into two parts. In the first part, we provide an in-depth explanation of the practice of evidence-based medicine, the problems associated with the practice, and explain how NLP can solve many of these problems. In the second part of our review, we explore automatic text summarisation techniques, starting from classic, domain-independent approaches up to recent approaches specific to the domain of our interest. We also discuss automatic summary evaluation techniques in this chapter.

Chapter 3 presents a detailed explanation of the data we use. In particular, it discusses our corpus, which is specialised for summarisation for evidence-based medicine. It also presents various corpus statistics, the annotation process which was carried out as part of this research, and examples from the corpus.

Chapter 4 describes our approach to the problem of automatic quality grading of evidence. The chapter presents our analysis of features, the rule-based approach for identifying the publication

types of articles automatically, and the combination of a number of useful features in a supervised machine learning task to predict evidence grades indicating the qualities of evidences.

Chapter 5 details our extractive summarisation approach. It presents our approach in collecting important statistics from the corpus, the use of target-sentence-specific scoring, the incorporation of domain knowledge in various forms, and the incorporation of question type information in our sentence scoring approach. It also provides the evaluation of our system relative to several well-established baselines.

In Chapter 6, we identify possible approaches for the generation of bottom-line summaries. We discuss our analysis of the corpus to validate the possibility of a two-step summarisation process. Finally, we explore two abstractive summary generation techniques, and discuss the use of context-sensitive polarity classification techniques to generate bottom-line recommendations.

Finally, in Chapter 7, we conclude with a summary of the thesis, outlining future directions and possible applications of the work.

2 Literature Review

2.1 Introduction

In this chapter, we review some of the literature that is relevant for the research described in this thesis. We commence the chapter by providing a detailed description of evidence-based medicine, how it is supposed to be practised, its evolution over time, its goals, the obstacles faced by practitioners, and a brief overview of how various Natural Language Processing (NLP) techniques can aid the practice. This part of the review illustrates the motivation behind specialised NLP research to support the various aspects of evidence-based medicine practice. Following this discussion, we focus on the main topic of research in this thesis: automatic text summarisation. We first provide a brief review of generic summarisation techniques, starting from early approaches to more recent ones. The discussion presents the state-of-the-art in summarisation/question-answering technologies in this domain, and points out the gaps in research that need to be filled to implement end-to-end systems. We then provide a relatively detailed discussion of summarisation approaches that have been applied to the medical domain, and analyse some recent summarisation approaches in detail. In particular, we attempt to answer the following questions in this chapter:

- What is evidence-based medicine and what are the major obstacles hindering evidence-based medicine practice?
- What are the characteristics of text in the medical domain?
- What tools and resources are available for text processing in the medical domain?
- How can NLP techniques help solve some of the problems associated with evidence-based medicine?
- What is automatic text summarisation and what is its relevance to evidence-based medicine?

- What is the current state-of-the-art in automatic text summarisation, particularly for the medical domain?
- How can automatic text summarisation approaches for evidence-based medicine be evaluated?

The rest of the chapter is organised as follows: we discuss the practice of evidence-based medicine in Section 2.2; in Section 2.3 we discuss the tools and resources available to practitioners to practise evidence-based medicine; in Section 2.4, we discuss some of the obstacles associated with evidence-based medicine, and provide an overview of the ways in which NLP can help solve some of the problems; we review some literature associated with automatic text summarisation in Section 2.5; in Section 2.6 we focus our review on question answering and summarisation techniques specific to the medical domain; in Section 2.7 we briefly review some evaluation techniques for automatic text summarisation; we conclude the chapter in Section 2.8 by summarising the key discoveries from the literature review.

2.2 Evidence-based Medicine

The term 'evidence-based medicine' was defined initially as *"a systematic approach to analyse published research as the basis of clinical decision making"* [Claridge and Fabian, 2005] by a group of researchers at McMaster's University. In practice, it involves much more than just analysing published papers, and a more concrete and widely accepted definition of evidence-based medicine was coined by Sackett et al. [1996] who explained it as *"the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients"*. Current clinical guidelines urge physicians to practise evidence-based medicine when providing care for their patients. As the definition suggests, and as will be elaborated later in this section, evidence-based medicine requires an approach that integrates the best external evidence with individual clinical expertise and patients' choice. Good practice of evidence-based medicine involves finding and appraising current medical evidence before making a decision. Good doctors use both individual clinical expertise and the best available external evidence, and neither alone is enough.

The goal of evidence-based medicine is to improve the quality of patient care, particularly in the long run, through the identification and promotion of practices that work, and the elimination of ineffective or harmful ones [Selvaraj et al., 2010]. This requires clinicians to be open-minded and to try new methods that are scientifically proven to be effective, and to discard old methods that are not [Crawford, 2007]. It must also be mentioned that whatever the evidence, values and preference judgements are implicit in every clinical decision. This means that according to the evidence-based medicine philosophy, clinical decisions, recommendations, and practice guidelines must not

only attend to the best available evidence, but also to the values and preferences of the informed patient. Values and preferences refer not only to the patients' perspectives, beliefs, expectations, and goals for life and health, but also to the processes individuals use to consider the available options and their relative benefits, harms, costs, and inconveniences [Montori and Guyatt, 2008]. This is a crucial point in the modern definition of evidence-based medicine [Ismach, 2004]. In essence, evidence-based medicine is rooted in five linked ideas:

1. Clinical decision should be based on the best available scientific evidence;
2. The clinical problem – rather than the habits or protocols – should determine the type of evidence to be sought;
3. Identifying the best evidence means using epidemiological and biostatistical ways of thinking;
4. Conclusions derived from identifying and critically appraising evidence are useful only if put into action in managing patients or making patient-oriented healthcare decisions; and
5. Performance should be constantly evaluated.

Greenhalgh [2006] points out the role of mathematics in evidence-based medicine by defining it as *"the use of mathematical estimates of the risk of benefit and harm, derived from high quality research on population samples, to inform clinical decision making in the diagnosis, investigation or management of individual patients"*. Numerous other definitions of evidence-based medicine exist, and this perhaps indicates that it is one of those concepts which cannot be defined coherently with a single sentence. The philosophy of evidence-based medicine offers some apparent advantages including an emphasis on learning about the patient's problems and making hidden assumptions explicit (*e.g.*, about the preferences of the patient) [Sackett et al., 2000]. A patient-oriented approach is key to the practice – combining the best patient-oriented evidence with patient-centred care, placing the evidence in perspective with the needs and desires of the patient [Slawson and Shaughnessy, 2005]. In the following subsections we provide a brief history of evidence-based medicine and describe how it is currently practised. In later sections, we discuss the problems associated with evidence-based medicine practice, the resources and tools available for text processing in this domain and the solutions that NLP can offer.

2.2.1 History of Evidence-based Medicine

Although a formal approach to evidence-based medicine is relatively new, the underlying concept has a long history. Claridge and Fabian [2005] provide a brief history of evidence-based medicine

and describe the various eras or phases during which the practice evolved. Evidence-based medicine in ancient times consisted mainly of historical and anecdotal accounts of disease treatments. Records of evidence-based treatments of diseases date as early as the Egyptian civilisation. However, the modern approach to evidence-based medicine did not commence until a few hundred years ago. Roughly during the seventeenth century, the *renaissance era* of evidence-based medicine began, and, during this time, personal journals and text books detailing disease treatments became more prominent. During the twentieth century, evidence-based medicine underwent a *transitional phase* as the sharing of information through textbooks and peer reviewed journals became more popular. This was followed by the *modern era* of evidence-based medicine, which began around the 1970's and was primarily driven by rapid technological advancements. Powerful computers and databases with massive storage capacity allowed the compilation of large amounts of data which could be used to perform evidence-based medicine. The emergence of the internet has given a massive boost to information sharing and consequently helped in the growth of evidence-based medicine by providing healthcare practitioners with access to a vast amount of data and information. However, the abundance of information has also given rise to problems in information retrieval and extraction, and solving these problems is an area of active research.

2.2.2 Evidence-based Medicine in Practice

Formally, the use of evidence in clinical decision making involves *"the practice of assessing the current problem in the light of the aggregated results of hundreds or thousands of comparable cases in a distant population sample, expressed in the language of probability and risk"* [Greenhalgh, 1999]. The practice, therefore, goes far beyond simply using the practitioner's clinical expertise, experience and knowledge of medical literature. It involves the efficient use of information search strategies to locate reliable and up-to-date information from varying sources and extraction strategies to efficiently collect and analyse retrieved information. The practice includes a process formally known as the *Critical Appraisal Exercise* which involves the following steps [Sackett et al., 1991, Selvaraj et al., 2010]:

1. defining a patient problem and the information that is required to resolve the patient's problem;

2. conducting an efficient literature search;

3. selecting the best of the relevant studies, and applying the rules of evidence-based medicine to determine their validity;

4. presenting the strengths and weaknesses of the evidence in an effective manner; and

5. extracting the relevant evidence and applying it to the patient care problem.

Practising evidence-based medicine involves applying these steps in real time. When a clinical encounter generates a question, it is researched on a real-time basis and immediately incorporated into the decision making process [Ismach, 2004]. Consequently, these are perhaps the most important steps in evidence-based medicine practice, and we now look at each of these in more detail.

Defining a Patient Problem as a Clinical Question

Formulating the patient problem forms the basis for the clinical question, which is used to search resources for an evidence-based answer. This is the first step in obtaining evidence for patient care. This is not an easy task and a well formulated question includes information about a patient (symptoms, signs, test results and knowledge of previous treatments), the particular values and preferences of the patient and other factors that could be relevant [Greenhalgh, 2006]. All that information should be summarised into a succinct question defining the problem and the specific additional items of information needed to solve the problem. Sackett et al. [2000] mention three important aspects of a good clinical question:

- First, define precisely *whom* the question is about.
- Next, define *which* manoeuvre is being considered in the patient or populations (*e.g.*, drug treatment), and, if necessary, a comparison manoeuvre.
- Finally, define the desired *outcome*.

Good question formulation can be best explained through an example. Consider the following two versions of the same question modified from Greenhalgh's [2006] examples:

- *Mrs X has developed light-headedness on these blood pressure tablets and she wants to stop all medication; what should she be advised to do?*
- *In a 68-year-old white woman with essential hypertension, no coexisting illness, and no significant past medical history, do the benefits of continuing therapy with bendrofluazide (chiefly, reduced risk of stroke) outweigh the inconvenience?*

Clearly, the second question provides more important details than the first one. It reveals important factors such as age, coexisting illnesses, medical history, and so on. Thus, while the first question

is almost unanswerable without using additional context, the second question provides sufficient information for a relevant literature search.

There has been substantial research in the area of medical question formulation and query-focused summarisation, and, in recent years, particularly in the field of evidence-based medicine (*e.g.*, a recent example of research drive in this area is the CLEF eHealth shared tasks [Suominen et al., 2013]). This is because it has been shown that the answerability of questions can be largely increased by better query formulation among other things [Gorman and Helfand, 1995]. The PICO format, which has four components, has become the accepted framework for formulating patient-specific clinical questions [Richardson et al., 1995]. The four components are: primary **P**roblem/**P**opulation, main **I**ntervention, main intervention **C**omparison, and **O**utcome of intervention. These components reflect key aspects of patient care, are recommended for the practice of evidence-based medicine and were originally developed for therapy questions, only to be extended later to all types of clinical questions [Armstrong, 1999]. Studies have shown that this framework improves the clarity of clinical problems and results in more precise search results [Booth et al., 2000, Cheng, 2004]. Although the PICO format has gained popularity with time, it is well known that not all clinical questions (even entirely evidence-based ones) can be mapped in terms of PICO elements, and this is particularly true for non-therapy questions [Huang et al., 2006]. There is also evidence that even doctors find it difficult to formulate the questions in terms of the PICO format [Ely et al., 2002]. Variants of the framework have been proposed (*e.g.*, PESICO [Scholosser et al., 2006], PIBOSO [Kim et al., 2011]); they offer more flexibility and comprehensiveness, and have applications beyond query formulation.

Conducting Literature Search

The medical domain has large amounts of lexical resources, and searching for relevant literature requires expertise in this area. An early study claimed that there were over 20 million medical articles[1] available, and that, every month, thousands of medical journals were published worldwide [Katz, 2001]. These articles are scattered over many databases, and obtaining the relevant lexical resource requires searching these databases. A number of databases, search engines and other tools are dedicated to providing access to medical literature, as we will see in the next subsection. Searching for appropriate literature can include searching from raw databases, databases with search filters, databases of pre-appraised articles, databases of synthesised articles and even personal contact with human sources. Some research has been carried out on strategies for retrieving high quality medical articles [Haynes et al., 1994, Hunt and McKibbon, 1997, Shonjania and Bero, 2001, Montori et al., 2005]. Searching for the correct literature is a rather tedious task, and it is in fact one of the major problems associated with evidence-based medicine

[1] The Medline database (discussed later) currently indexes over 22 million medical articles.

practice [Ely et al., 2005].

Selecting the Best Resources

The selected articles must be closely relevant to the problem at hand, and, at the same time, they must have a high 'level of evidence'. The 'level of evidence' of a medical publication may depend on a number of factors such as the publication type and the number of subjects involved in the study. Checking the relevance of the papers requires a thorough analysis, and a ranking system is usually employed for examining the 'level of evidence' of different sources. Medical publication types include Systematic Reviews (SR), Randomised Controlled Trials (RCT), Meta-Analyses (MA), single Case Studies, Tutorial Reviews and many more, including even personal opinions. Although all of them provide evidence of some form, their levels of evidence vary significantly. Figure 4.3, slightly modified from the one provided by Gilbody [1996], provides a ranking of some of the common medical article types (highest ranked on top).

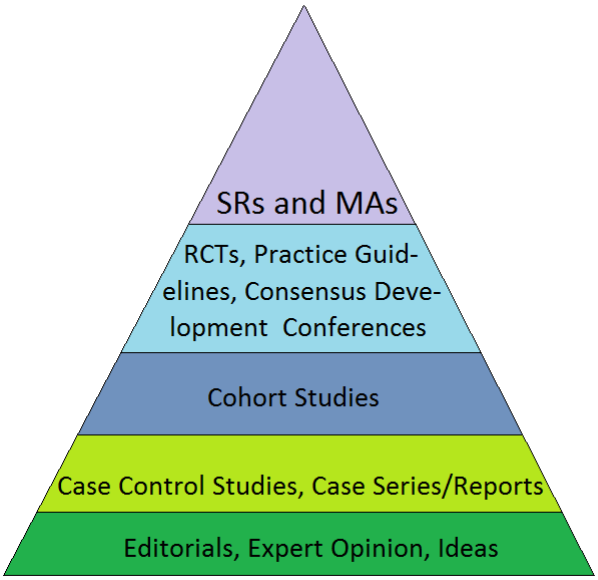


Figure 2.1: Quality of evidence with respect to publication types, adapted from Gilbody [1996].

A detailed analysis of these different types of studies is outside the scope of this survey. Guidelines are available for healthcare practitioners to obtain necessary information from each of these types of studies and evaluate their levels of evidence. For example, one of the earliest guidelines is provided by Sackett et al. [1991]. It provides clear and concise criteria for Randomised Controlled Trials, Case Control Studies and so on. To date, there are numerous books and other resources available, providing the required guidelines for selecting articles while practising evidence-based medicine. Examples of guidelines available online include ASSERT (A Standard for the Scientific

and Ethical Review of Trials)[2], PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)[3] and EQUATOR (Enhancing the QUAlity and Transparency Of health Research)[4].

Identifying Strengths and Weaknesses

Once relevant papers are identified, each paper must be studied in detail to extract the evidence it contains with respect to the problem at hand. Practitioners are particularly interested in studying in detail what type of study was conducted, on how many subjects, where the subjects came from, what treatment or other intervention was offered, how long the follow up period was and what outcome measures were used [Greenhalgh, 2006]. For quantitative tests, analysing the results of statistical tests is also an important task. To the practitioner, the single most important detail contained in each article is the explanation of the results, together with information about their reliability. Thorough analysis of these reveals the strengths of the studies and exposes their weaknesses or shortcomings. Identification of the strengths and weaknesses provides a clear indication of the quality of the evidence provided by each article, and hence, aids the practitioner to make the final decision.

Applying Evidence to Patient Care

Practitioners make their final judgement considering the outcomes presented in the article(s) and the relevance of the article(s) to the problem at hand. Often, a number of articles suggest the same solution or a similar one, making it easier for the practitioner to make a decision. However, there are also cases when chosen articles provide contradictory outcomes. In such cases, practitioners have to choose between outcomes based on personal experience, the 'quality of evidence' of the articles, the closeness of the articles to the given problem or other sources of evidence. Note that the 'quality of evidence' here refers to the overall quality of the evidence obtained from all the articles combined, whereas the 'level of evidence' of a single article refers to the reliability of that particular article only. The following subsection further elaborates on this topic.

Generally, countries have mandatory clinical practice guidelines that must be adhered to when performing evidence-based medicine. Practitioners primarily navigate through these guidelines when practising evidence-based medicine. As such, it is actually often the providers/developers of clinical guidelines who are required to perform elaborate literature searches and follow the approaches mentioned above for the preparation of the appropriate practice guidelines. Thus, the primary users of the system we describe in this thesis are not only healthcare practitioners, but

[2] http://www.assert-statement.org. Accessed on 26th May, 2014.
[3] http://prisma-statement.org. Accessed on 26th May, 2014.
[4] http://www.equator-network.org. Accessed on 26th May, 2014.

also, crucially, the developers of clinical practice guidelines.

2.2.3 Quality of Evidence

The quality, strength or grade[5] of a recommendation for clinical practice is based on a body of evidence typically consisting of more than one study. This usually takes into account the level of evidence of the individual studies; the type of outcomes measured by these studies; the number, consistency and coherence of the evidence as a whole; and the relationship between benefits, harms and costs [Ebell et al., 2004]. Various organisations and publications have their own measure of evidence and, according to a research report [West et al., 2002] produced by the Agency of Healthcare Research and Quality (AHRQ), more than 100 grading scales are in use today. The report also proposes that any system for grading the strength of recommendation should consider three key elements: Quality (the extent to which the identified studies minimise the opportunity for bias), Quantity (the number of studies and subjects included in those studies) and Consistency (the extent to which findings are similar between different studies on the same topic).

There has been research to identify a grading system that is suitable for the practice of evidence-based medicine. Among other requirements, studies have specified the need for a balance between simplicity (such that assessing the quality of evidence is not very time-consuming) and clarity (so that evidence can be easily classified into a specific grade) [Atkins et al., 2004]. Comprehensiveness of grading systems is also seen as an important factor [Ebell et al., 2004] since they need to be applied to studies of screening, diagnosis, prevention, therapy and prognosis. Despite the presence of a variety of grading systems, only a minority of them adequately address these requirements. Carrying out an in-depth review of all the popular grading systems is outside the scope of this paper. Therefore, we focus, for the rest of this subsection, on the Strength of Recommendation Taxonomy (SORT) grading scale, which is the grading system we use in this thesis.

The SORT Grading Scale

SORT was first proposed in 2004 through a collaborative effort by the editors of multiple family medicine journals with the purpose of providing authors and readers of family medicine journals with a simple, user-friendly system for the grading of evidence. SORT provides a uniform recommendation-rating system that can be applied throughout the medicine literature. Its simplicity and straightforwardness increases its usefulness to practitioners [Weiss, 2004].

[5]The three words – quality, strength and grade – are used synonymously in this thesis when referring to an evidence-based recommendation.

This taxonomy uses only three ratings – **A** (strong), **B** (moderate) and **C** (weak) – to specify the strength of recommendation of a body of evidence [Ebell et al., 2004]. Grade **A** reflects a recommendation based on *consistent* and *good-quality* [a good-quality evidence consists of high quality Systematic Reviews, Meta-Analyses or Cohort Studies with good followup], *patient-oriented* evidence. Grade **B** reflects a recommendation based on *inconsistent* or *limited-quality*, *patient-oriented* evidence. Grade **C** reflects a recommendation based on consensus, usual practice, opinion, *disease-oriented* evidence, or Case Series for studies of diagnosis, treatment, prevention or screening. An important aspect of SORT is that it allows grade conversion to other popular grading scales, such as the ones used by the Centre for Evidence-Based Medicine (CEBM)[6] and the BMJ Publishing Group, enabling authors, editors and readers to move between taxonomies [Ebell et al., 2004].

Chapter 4 of this thesis focuses on the automatic assessment of the quality of medical evidence using the SORT scale. Details about this research work, and further details about the SORT system are provided in that chapter. We use SORT as our target grading scale because of its flexibility and also because of the availability of annotated data based on this grading scale.

2.3 Resources for Evidence-based Medicine and the Medical Domain

Evidence-based medicine requires practitioners to stay up-to-date with the latest medical literature and use the latest research discoveries. This is a daunting task. There are resources available on which the healthcare practitioner can rely. In this section, we specify some of the resources and tools available for this practice.

2.3.1 The Medline Database and Similar Databases

The Medline database is maintained by the National Library of Medicine (NLM), U.S.A., and it is the most popular source of up-to-date evidence [Taylor et al., 2003]. It indexes over 5,000 journals published in over 70 countries and holds more than 22 million records. Medline is available online, either via the NLM PubMed[7] interface or from commercial vendors who use their own search engines. PubMed incorporates a lexical resource of medical terms, the Medical Subject Headings (MeSH)[8]. There are 26,853 descriptors in the 2013 version of MeSH. There are a total of 213,00 entry terms that assist in finding the most appropriate MeSH headings. Queries to PubMed are analysed and expanded with MeSH terms, thus increasing the likelihood

[6] http://www.cebm.net/?o=5653. Accessed on 26th May, 2014.
[7] http://www.ncbi.nlm.nih.gov/pubmed. Accessed on 26th May, 2014.
[8] http://www.nlm.nih.gov/pubs/factsheets/mesh.html. Accessed on 26th May, 2014.

of finding documents that use variations of the same term. Furthermore, the articles returned contain meta-data in addition to the MeSH terms, some of which can be useful to determine the quality of evidence. For example, skilled searchers may specify the publication types of articles to obtain better quality articles. The database therefore attempts to aid search and retrieval of relevant articles. Information Retrieval (IR) research in this area relies largely on the extra information available with each article for efficient and accurate retrieval.

Despite the broad coverage of Medline and its impressive collection, many important medical and paramedical journals are not covered by this database, particularly journals not published in the United States. Hence practitioners often have to refer to other similar databases that specialise in the required areas. The following is a list of databases that are relevant to various areas. Note that providing details of each database is outside the scope of this thesis, and therefore appropriate links to the databases are provided as footnotes. The list of databases is by no means exhaustive.

- Allied and Complementary Medicine (AMED)[9] – Produced by the British Library, it covers a range of complementary and alternative medicine including homeopathy, chiropractic, acupuncture, and so on.
- CINAHL[10] – It covers all aspects of nursing, health education, occupational therapy, social services in health care and other related disciplines.
- Current Contents Search[11] – Updated weekly, it indexes journal tables of content on or before their publication date.
- Embase[12] – It is the database of Excerpta Medica which focuses on drugs and pharmacology, clinical medicine and other biomedical specialities. It is more up-to-date than Medline, with more detailed indexing and better European coverage.
- PsycInfo[13] – Produced by the American Psychological Association, it covers psychology, psychiatry and related subjects.

2.3.2 Databases of Synthesised Information

One problem with Medline and other 'raw' databases is that they contain articles of varying quality – from high quality SRs and MAs to informal, unreliable clinical trials. When specific

[9] http://www.ebscohost.com/academic/AMED-The-Allied-and-Complementary-Medicine-Database. Accessed on 26th May, 2014.

[10] http://www.cinahl.com. Accessed on 26th May, 2014.

[11] http://thomsonreuters.com/current-contents-connect/. Accessed on 26th May, 2014.

[12] Available from a number of database suppliers including Ovid (http://www.ovid.com. Accessed on 26th May, 2014.)

[13] Available from a number of database suppliers including Ovid (http://www.ovid.com. Accessed on 26th May, 2014.)

topics are searched for in these databases, the results returned invariably contain a mixture of high and low quality articles. Practitioners, particularly those practising evidence-based medicine seek high quality articles, and therefore have to manually appraise the search results of raw databases, which is understandably time-consuming. To address this problem, there are databases that only contain articles that are of high quality and are composed by synthesising multiple high quality articles. A good example of such a database with synthesised evidence is the Cochrane Library[14]. It contains over 4,000 peer-reviewed Cochrane Systematic Reviews, over 5,000 Systematic Reviews listed in the Database of Abstracts of REviews (DARE), and about a half-million selected published clinical trials in their Central Registry of Controlled Trials. However, the number of articles contained in this library is minute compared to the total number of medical articles available, and the scope of topics covered is quite limited [Silagy et al., 1999]. Despite this, the Cochrane Library is very much the first port of call for clinical researchers looking for quality articles.

Among such high quality databases, there are also guideline databases such as the National Guideline Clearing House[15] and National Institute for Health and Clinical Excellence (NICE)[16] that provide evidence-based clinical practice guidelines. Many countries have their own national clinical practice guidelines, and sometimes their use are mandated by law (*e.g.*, in Finland). There are also sources that not only synthesise the best available information but also present them in readily usable formats. The following is a brief, non exhaustive list of such sources:

- Clinical Evidence (CE)[17] – Produced in the United Kindom, it synthesises best evidence in mainstream clinical areas and is published in a book format every six months.
- PIER – It is a US database equivalent to CE maintained by the American College of Physicians.
- Evidence-Based On-Call (EBOC)[18] – It is a database that contains evidence-based summaries for specific conditions and is designed for junior doctors on call.

2.3.3 Other Sources

Practitioners often rely on other sources of information when practising evidence-based medicine. There are databases containing pre-appraised articles, for example, which are generally fairly small databases listing clinical research papers that have already been manually appraised.

[14] http://www.cochrane.org. Accessed on 26th May, 2014.
[15] http://www.guideline.gov. Accessed on 26th May, 2014.
[16] www.nice.org.uk. Accessed on 26th May, 2014.
[17] http://www.clinicalevidence.com/ceweb/conditions/index.jsp. Accessed on 26th May, 2014.
[18] http://www.eboncall.org. Accessed on 26th May, 2014.

Journals such as ACP Journal Club, Journal of Family Practice, Evidence Based Medicine and Evidence Based Mental Health are examples of such sources. Although their coverages are relatively narrow, the information contained in these sources are very reliable. The Journal of Family Practice[19] has a “Clinical Inquiries" section that contains clinical questions and evidence based answers to the questions along with the strength of recommendations of the answers using the Strength of Recommendation Taxonomy. With time, the contents of these sources of information are increasing, and they are becoming more useful as resources for evidence-based medicine.

Another resource that has become popular in recent years is UpToDate[20] which covers over 8,300 topics in 16 medical specialities and includes more than 97,000 pages of text, graphics, links to Medline abstracts, more than 385,000 references and a drug database. The UpToDate community includes over 5,100 expert clinicians who review and update content continuously. UpToDate is evidence-based and uses a literature-driven updating system; more than 440 journals are monitored by editors and authors, and, anytime something of importance is published, it is incorporated into the resource.

In recent years, progress in IR research has greatly improved the performance of search engines across all domains. There are a number of search engines specialised for the medical domain, and about.com[21] provides an analysis of the top five search engines in the medical domain. The list includes PubMed, OmniMedicalSearch[22], WebMd[23], Healthline[24] and HealthFinder[25].

Generic search engines are also frequently used to search for evidence and they, particularly Google, have evolved into effective search tools for online full text peer reviewed journals [Greenhalgh, 2006]. According to a study by Tutos and Mollá [2010] on the use of search engines for answering clinical questions, Google performs better than any other systems including PubMed which is specialised on medical text. This clearly demonstrates the strength of generic search engines and their emergence as a useful tool in medical article retrieval and evidence-based medicine practice.

[19] http://www.jfponline.com. Accessed on 26th May, 2014.
[20] http://uptodate.com/home/about/index.html. Accessed on 26th May, 2014.
[21] http://websearch.about.com/od/enginesanddirectories/tp/medical.htm. Accessed on 26th May, 2014.
[22] http://www.omnimedicalsearch.com. Accessed on 26th May, 2014.
[23] http://www.webmd.com. Accessed on 26th May, 2014.
[24] http://www.healthline.com. Accessed on 26th May, 2014.
[25] http://www.healthfinder.gov. Accessed on 26th May, 2014.

2.4 Problems Associated with Evidence-based Medicine

In this subsection, we briefly review and analyse the problems and barriers faced by evidence-based medicine practitioners. This review expresses the technological needs of evidence-based medicine and introduces the possibilities presented by NLP in this area.

Evidence-based medicine practice requires obtaining relevant clinical information from medical literature and combining the best available evidence. Most problems associated with evidence-based medicine emerge from this requirement. As already mentioned, there is an abundance of information available to healthcare practitioners, resulting in practitioners being overwhelmed by the amount of information available to the point that they often cannot obtain evidence-based answers to their questions about specific clinical problems [Covell et al., 1985, Williamson et al., 1989, Ely et al., 1999, Coumou and Meijman, 2006].

There has been a number of studies regarding the problems associated with evidence-based medicine practice. Ely et al. [2002] conducted a well known research to identify and describe the obstacles practitioners face when attempting to answer clinical questions with evidence. The study revealed fifty-nine obstacles which were divided into the following five broad categories:

1. Recognise a gap in knowledge – Practitioners are often not aware of their lack of knowledge about specific topics and attempt to answer clinical questions based on their own experience only. Also, practitioners aware of a gap in knowledge sometimes decide to ignore it due to various reasons such as time pressure.

2. Formulate a question – Questions formulated by practitioners often lack necessary information, as already discussed earlier in this chapter, and therefore retrieved results are either incomplete or not fully relevant.

3. Search for relevant information – Due to the overwhelming amount of information, practitioners often do not know where to search for relevant information or how to search for it in a limited amount of time.

4. Formulate an answer – Often articles on specific topics provide information without completely answering the clinical question posed by the practitioner, making answer formulation difficult. Studies have shown that practitioners prefer receiving summarised answers rather than searching and formulating answers themselves [Barry et al., 2001].

5. Use the answer to direct patient care – When applying evidence for patient care, the patients' preferences, medical history, and other related factors must be taken into account, and this often poses a problem. For example, a patient's choice may conflict with the

best clinical evidence, and such conflicts must be resolved by practitioners [Sanders et al., 2008].

From the fifty nine obstacles, Ely et al. [2002] also identified six that are considered particularly salient by practising doctors:

1. the excessive time required to find information;
2. the difficulty to modify the original question, which is often vague and open to interpretation;
3. the difficulty to select an optimal strategy to search for information;
4. the failure of a seemingly appropriate resource to cover the topic;
5. the uncertainty about how to know when all the relevant evidence has been found so that the search can stop; and
6. the inadequate synthesis of multiple bits of evidence into a clinically useful statement.

We now look at these obstacles in more detail.

2.4.1 Problems Associated with Time

The time associated with seeking information is largely considered to be the biggest obstacle in evidence-based medicine practice [Verhoeven et al., 2000, Smith, 1996, Westberg and Miller, 1999, Dorsch, 2000, McColl et al., 1998, Wilson, 1999, Ely et al., 2002, Coumou and Meijman, 2006]. Due to the time-consuming nature of the practice, it has been argued that this approach to patient care ignores patient values [Graham and Grondin, 2007]. Although the skills of searching for evidence and critically appraising it are being mastered by growing numbers of doctors, many cannot keep up. Consequently, there is a widening chasm between what is ought to be done and what is actually done [Davidoff et al., 1995]. Most busy doctors lack the time or skills to track down and evaluate evidence, and when searching for evidence, practicing doctors do not have time to search multiple sites or scroll through long bodies of texts. Nor do they have time to search multiple textbooks or perform literature searches for most of their questions. They need to pick the right resource the first time, the information in that resource needs to be readily found, and all the information must be there [Ely et al., 2002]. Literature search and appraisal may take hours and even days. According to Hersh et al. [2002], it takes more than 30 minutes on average for a practitioner to search for an answer. But usually practitioners spend about 2 minutes [Ely et al., 2000]. Hence, many questions go unanswered.

Ely et al. [2005] classify this problem as a 'resource-related' problem and state that physicians want rapid access to concise answers that are easy to find and tell them what to do in specific terms. Physicians have many *ad-hoc* clinical questions at the moment of patient care but limited time and resources to search for answers [Yu and Cao, 2008]. Ismach [2004] elaborates on this issue and explains that, in most cases, such as in the emergency department, time is rarely available for elaborate or comprehensive literature search. In a lot of cases, practitioners choose not to pursue answers to clinical questions because of the lack of time available [Ely et al., 2005], and attempt alternative solutions instead.

2.4.2 Problems Associated with Question Formulation

This is a physician-related problem and is a consequence of the tendency of physicians to formulate questions in a way that is difficult to answer from general resources [Ely et al., 2005]. Patient-specific questions tend to be vague (*e.g.*, 'what is this rash?') [Ely et al., 2002] and therefore cannot be answered by general resources. Physicians typically ask difficult questions and hence only pursue evidence for them in a minority of cases [Ely et al., 1999]. The PICO format, discussed earlier, is sometimes used to aid practitioners formulate their questions. Recent research in medical information retrieval has focused on query formulation and other aspects of information retrieval to aid practitioners [Heppin and Jarvelin, 2012, Kelly et al., 2014].

2.4.3 Problems Associated with Search Strategy

The medical literature has been described as 'unwieldy, disorganised and biased' [Godlee, 1998], and searching for specific topics in the vast amount of available information is analogous to searching for a needle in a haystack. Although electronic databases facilitate searching, the procedure requires skill and expertise, and, even for an experienced librarian, the search to find a comprehensive set of documents related to a focused clinical query may take hours. The first obstacle that doctors face is the uncertainty of where to look for information [Ely et al., 2002]. Other difficulties related to information seeking include: the presence of a large number of irrelevant material in search results, the difficulty in finding correct search terms, inefficient indexes in books and journals and badly organised journal volumes [Verhoeven et al., 2000]. Furthermore, it can be difficult to decide which resources will be most helpful and what should determine the selection of resources. Also, the search strategy must conform to the amount of time available, the practitioner's familiarity with the resources, and the type of question.

Lack of expertise in information searching by practitioners has also been mentioned in the literature as a major obstacle, and practitioners have often expressed the need for training on this subject [Wilson, 1999, Wilson et al., 2001]. Studies have also shown that practitioners' access to

a computer and the internet varies largely with their location, and this significantly affects the information search strategy [Wilson et al., 2001, Kalsman and Acosta, 2000, McColl et al., 1998, Williams and Maj, 2001].

2.4.4 Problems Associated with Source of Evidence

Often practitioners do not have access to the necessary resources required to answer a clinical question [Young and Ward, 2001]. One of the most common obstacles encountered by practitioners when pursuing a clinical question is not finding the needed information in the source selected [Ely et al., 2005]. Hence, often practitioners are simply unable to find appropriate resources that can answer their clinical questions. Even for practitioners with adequate access to medical literature, it is often impossible to find sources of information that fully answer their questions due to the complexity of clinical questions. Ely et al. [2002] also point out that since general resources do not allow real time interaction with the searcher, it is not possible for them to use follow up questions to refine searches.

2.4.5 Problems Associated with the Relevance of Evidence

Looking for evidence for a particular question requires a practitioner to shortlist a set of papers that seem relevant after an initial search, and then study the papers in detail [Greenhalgh, 2006]. However, the number of relevant documents can vary significantly between topics, and, therefore, physicians are often left unsure as to when they should stop searching [Ely et al., 2002, Green and Ruff, 2005]. This is a significant obstacle because often practitioners spend time unnecessarily searching for evidence when the required evidence has already been found. Alternatively, practitioners may also rely on incomplete evidence not knowing that better evidence exists. Furthermore, evidence may be directed at the wrong audience (*e.g.*, patients) and therefore may not be relevant to doctors [Ely et al., 2002].

2.4.6 Problems Associated with Information Synthesis

Once all the relevant information has been found, it is quite a daunting task to synthesise the information from different sources. Studies carried out have shown that practitioners frequently mention difficulties with generalising research findings and applying evidence to individual patient care [Young and Ward, 2001]. The reasons behind this include the incapability of any of the sources to completely answer the clinical question, the selected articles not directly answering the clinical question, and different articles providing contradictory information [Ely et al., 2005, 2002].

2.4.7 Other Obstacles to Evidence-based Medicine

There are numerous other obstacles to evidence-based medicine in addition to the ones already mentioned. It has been shown that almost half the questions asked by practitioners are not pursued at all for various reasons [Ely et al., 2005]. Common reasons include the expectation that no useful information would be found and the tendency to consult colleagues rather than performing an actual search [Coumou and Meijman, 2006]. Often, when a question is pursued, the results returned are not from reliable sources and therefore cannot be used by the practitioner.

2.4.8 Possible Applications of IR and NLP in Evidence-based Medicine

Despite the presence of many barriers, evidence-based medicine practice has gained popularity over recent years for a number of reasons, including its promise of improving patient healthcare in the long run. As for the barriers, advances in technology are gradually eliminating them and making the process more efficient. As mentioned earlier in the chapter, the modern approach towards evidence-based medicine has been made possible through the presence of electronic databases and is driven primarily by connectivity via the internet. From the problems and barriers specified earlier in this section, it can be inferred that the next boost in evidence-based medicine practice will come from research in NLP.

NLP offers suitable solutions to the problems faced by evidence-based medicine practitioners, particularly for problems associated with information overload. Practitioners require comprehensive, specific bottom-line recommendations that anticipate and directly answer clinical questions. They require rapid access, current information and evidence based rationale for recommendations [Ely et al., 2005]. NLP has the potential of addressing all these requirements. For example, **Query Analysis** can be used to understand and expand practitioners' queries through the use of domain-specific semantic information. Queries posed by practitioners are often very short with an average of 2.5 words [Hoogendam et al., 2008], and therefore existing techniques for ontology-driven query reformulation [Schwitter, 2010] can be built upon to help practitioners compose and refine queries. **Information Retrieval** techniques tailored for the medical domain can be used to increase the recall and precision of literature searches. Strength of recommendation values can be used to classify documents to make searches more reliable. There has already been some research in this area with promising outcomes [Karimi et al., 2009, Pohl et al., 2010]. **Information Extraction** techniques which incorporate domain knowledge (*e.g.*, MeSH terms, UMLS) can be used to extract relevant information, based on practitioners' questions, from the retrieved documents. This is also an area that is being explored by medical informatics researchers, and knowledge-based and statistical techniques have produced promising results [Lin and Demner-Fushman, 2007]. Furthermore, some research has attempted to extract semantic information and use that information to represent the main concepts presented in documents [Fiszman et al., 2003,

2004, 2009]. However, the results obtained using these techniques are still not accurate enough to be used in fully automatic clinical decision support systems. Research in the area of information extraction is closely related to that of **Document Summarisation**. The goal of summarisation in this context is to summarise the content extracted from multiple documents and present them to the users (i.e., specific bottom-line recommendations). Summarisation of multiple documents is a well explored area (*e.g.*, in the news domain), however its application to the medical domain is still quite limited. Successful summarisation of medical documents and effective presentation of summarised information to the user are key to the successful development of end-to-end question answering systems that can be used for evidence-based medicine practice. More specifically, query-focused multi-document summarisation that incorporates medical domain knowledge is an area of research that is worth exploring and success in this area will significantly advance the practice of evidence-based medicine.

Our research on automatic evidence-based summarisation is motivated by these factors. In the following subsections of this chapter, we review the literature associated with text summarisation, particularly in the medical domain.

2.5 Automatic Text Summarisation

Before we dive deep into summarisation approaches specific to the medical domain, it is important to discuss the basics of text summarisation and look at some approaches to summarisation already proposed. We attempt to achieve this goal in this section. It must be mentioned, however, that this is by no means a complete survey of automatic text summarisation, since the field of automatic text summarisation is too large to be discussed in a single survey. Instead, we primarily discuss important domain-independent techniques and breakthroughs in this area, and approaches that we utilise in the research work described in this thesis. We have attempted to keep this section short with references to detailed literature for the interested reader.

2.5.1 Overview

According to Radev et al. [2002], the goal of a summary is to present the main ideas of a document in less space. Mani [2001] provides a more formal definition and explains that the process of summarisation involves *taking an information source, extracting content from it, and presenting the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs*. It also involves distinguishing between more and less informative segments in a document and choosing the informative segments at the expense of the rest of the documents. Sparck Jones [1999] explains that summarisation is a hard task because it requires characterisation of a source text as a whole, capturing its important content, where content is a

matter of both information and its expression, and importance is a matter of what is essential as well as what is salient. Das and Martins [2007] present three important aspects that characterise research on text summarisation:

- Summaries may be produced from a single document or multiple documents,
- Summaries should preserve important information, and
- Summaries should be short.

The motivation for building automatic summarisation systems has increased over time due to the increasing availability of web-based textual information, and the explosion of available information has necessitated intensive research in this area [Das and Martins, 2007, Fattah and Ren, 2009, Mani et al., 2000]. The number of online text documents has been increasing exponentially over the recent years, and this is also true for the medical domain [Hunter and Cohen, 2006]. Consequently, significant progress in automatic summarisation has been made in the last decade [Sparck Jones, 2007].

At this point, it is perhaps wise to briefly present the pros and cons of having an abundance of textual information available. In short, the prime advantage of having a large amount of available information is the redundancy in it. Research has shown how summarisation systems can exploit this redundancy [Barzilay and McKeown, 2005, Clarke et al., 2001, Dumais et al., 2002], particularly when summarising from multiple documents. As will be discussed later, many summarisation systems rely on redundancy (*e.g.*, frequently occurring words, concepts, etc.) to generate automatic summaries. In evidence-based medicine, redundancy is beneficial as it gives stronger evidence on a given finding. As for the cons, the abundance of information also introduces the need for efficient information retrieval and extraction, both of which are difficult tasks. In many cases, identification of relevant information requires elaborate manual searching through redundant information, which is often quite time consuming and therefore rather inefficient [Barzilay and McKeown, 2005]. It has been realised that added value is not gained simply through larger quantities of data, but through easier access to the required information at the right time and in the most suitable form [Afantenos et al., 2005]. This gives rise to the need for technologies that can gather required information for the users and present them in a simplified, concise and friendly manner. Thus, we can say that while the large amount of information is a necessary condition for the development of automatic summarisation systems, it also introduces the problem of information overload.

In the rest of this section, we provide a brief description of the basic concepts associated with text summarisation followed by a review of some of the historical and recent approaches to automatic summarisation.

2.5.2 Factors of Text Summarisation

Automatic text summarisers must take into account a number of factors to achieve their goals. Here we discuss some of these factors, knowledge of which is essential to understand the process of text summarisation. The factors affecting automatic text summarisation can be grouped into three main categories: input, purpose and output. We primarily draw the following information from existing literature [Sparck Jones, 1999, Mani, 2001, Afantenos et al., 2005, Sparck Jones, 2007], and provide references to articles containing more detail in specific cases.

Input

The following factors are associated with the inputs of a summarisation system:

Unit - Summarisers can either take as input a single document or multiple documents. Early research on automatic summarisation focused mainly on single documents, but the availability of large collections of electronic documents on the same topics and early research in news summarisation led to the growing interest in multi-document text summarisation. While in single-document summarisation the summary consists of information extracted from one document, multi-document summarisation attempts to produce a single summary from a set of related documents [Radev et al., 2002]. Summarising multiple documents is a more difficult task than summarising single documents, and, additional algorithms are required to overcome problems of redundancy (different documents might present identical information which could lead to repetitions in the summary), inconsistency (the information presented in distinct documents may not be consistent), and incoherence (information extracted from separate documents may not be coherent when presented together). In the research work described in this thesis, we explore both single- and multi-document summarisation.

Language - A summarising system can be mono-lingual, multi-lingual or cross-lingual. In mono-lingual systems, the input and output are in the same language. Multi-lingual systems are capable of handling more than one language. In cross-lingual systems, the input and output languages are different. Our summarisation research focuses on English text only, so it is essentially a mono-lingual system.

Domain - Summarisation systems can either be domain-specific or domain-independent. Domain-independent approaches can be applied to documents from various domains without requiring any changes in the algorithm. While having the benefit of portability to different domains, domain-independent approaches fail to take advantage of knowledge and resources available to specific domains. Domain-specific systems, on the other hand, are designed for specific domains and use all the available resources and knowledge available for the relevant domain. There has

been significant research in text summarisation specific to domains such as news [Barzilay and McKeown, 2005, Christel et al., 2002, Papernick and Hauptmann, 2005, McKeown et al., 2002], scientific [Plaza et al., 2011a], legal [Moens et al., 1997, Farzindar and Lapalme, 2004], and medical [McKeown et al., 1998, Lin and Demner-Fushman, 2007, Rindflesch et al., 2005].

Text summarisation tasks in restricted domains may also include additional subtasks, depending on the needs of the end user. For example, news summarisation systems often need to take into account the dates and times of publication of the articles, so that the summaries can be presented in chronological order. For the research described in this thesis, we show that an essential sub-task of the overall summarisation task in the evidence-based medicine domain is to automatically assess the qualities of the medical evidences (described in Chapters 3 and 4).

Medium - Other than text, input can consist of speech, images, tables, and so on. There has been some research in the production of combined summaries for sources containing language and image [Christel et al., 2002, Papernick and Hauptmann, 2005]. A lot of research has also focused on producing text summaries from non-textual material [Jordan et al., 2004, Maybury, 1995, McKeown et al., 1995, Yu et al., 2007b].

Structure - This refers to the external structure of documents, such as headings, boxes, tables, and also structure embedded in text like familiar rhetorical patterns. The structure of documents can vary between domains. For example, news stories have little explicit structure other than top headings [Sparck Jones, 2007], while technical articles contain more structure which can be exploited by summarisation systems [Elhadad and McKeown, 2001, McKeown et al., 1998, Saggion and Lapalme, 2002, Lin and Demner-Fushman, 2007]. Medical article abstracts may be structured or unstructured, and this factor plays an important role in the performance of summarisers.

Meta-data - In some cases, header information or meta-data associated with input documents play an important role in summarisation. For example, dates of publications are vital for news items. In the case of the medical domain, databases such as Medline use meta-data to indicate the key topics in each article. We utilise the meta-data associated with the Medline entries in some of our tasks.

Other Input Factors - Source of documents, genre (*e.g.*, report, biography, description, etc.) and Authorship are some of the other input factors mentioned in literature that may affect summarisation systems.

Purpose

These factors are associated with the purpose of the summaries. In other words, these factors determine what a summary is for or what a summary is like.

Summary Type - Summaries can either be extractive or abstractive. The task of extractive summarisation is the easier of the two. It involves extracting content from the source document(s) and presenting them as the summary. The extracted content can be words, phrases, sentences or even paragraphs. It is therefore a process of information extraction. Abstractive summarisation, in contrast, involves identifying the most salient concepts prevalent in the source document(s), fusing the concepts and presenting them (usually through natural language generation techniques) [Afantenos et al., 2005].

Information - Summaries can be generic or user-oriented. Generic summaries only take into account the information found in the input document(s), while user-oriented summaries attempt to extract and summarise only the information that is relevant to a user's query. The query enables the user to formulate the information needs. Thus, query-oriented summarisation systems are user-focused, adapting each time to the expressed needs of the users, as viewed through the query they make [Afantenos et al., 2005]. Query-focused summaries, therefore, have to identify information needs posed by user queries.

Use - This is the most important purpose factor [Sparck Jones, 2007] and defines the purpose of a summary. This is the major influence on summary content and presentation [Sparck Jones, 2001]. Even relatively poor summaries produced by a system can be acceptable as long as the summaries fit their purpose (*e.g.*, they present the information requested by a query). Existing literature rarely refers to summary use; however, evaluation effort in this research area is now moving towards purpose-focused evaluation as a lever in system design and summary assessment.

Audience - Summaries must take into account the intended audience. Different summaries can be generated from the same input sources to suit the requirements of the audience. News material, for example, has at least two audiences: ordinary readers and analysts [Sparck Jones, 2007]. Elhadad [2006] produces two kinds of summaries for medical documents: one intended for healthcare practitioners and thus containing technical terms, and the other for non-medical personnel or 'laymen' and containing minimal medical terms.

Time - Time is a critical factor for some summarisation systems. For example, traffic alerts require timely delivery [Evans et al., 1995], and news summaries have to roll forward with developing story lines [McKeown et al., 2002]. Furthermore, query-oriented summaries need to be delivered promptly, and thus, corresponding systems need to be very fast. Ismach [2004], for example, mentions the need for fast information delivery in cases of medical emergencies.

Location - Location refers to the devices on which the automatic summaries will be displayed (*e.g.*, a desktop computer). This is important because this determines features such as the length of the summary, its content and visualisation.

Other Purpose Factors - Other purpose factors have been identified in the literature including formality (*e.g.*, legal constraints on the summarised information, author attributions etc.), the triggers for a summary (i.e., what causes the summary to be created: a user query or some other trigger) and destination (i.e., the intended recipient of the summary: can be humans or some other systems).

Output

These factors are associated with the summarised output.

Coverage - Coverage of a set of sources by a summary can either be comprehensive or selective. Reflective summaries are comprehensive (i.e., summarises the whole source text), while query- or topic-oriented summaries are selective (i.e., summarises only a portion of the source text that is relevant to the query/topic) [Sparck Jones, 2007].

Compression - This refers to the amount by which the information from the source(s) are reduced during summarisation. Compression may be influenced by the requirement of the user [Lin and Demner-Fushman, 2007] or the limitations imposed by the location (*e.g.*, handheld devices [Boguraev et al., 2001]).

Structure - The output of a summary may be plain text or the information may be represented using tables (*e.g.*, Mani et al. [2000]), forms (, Farzindar and Lapalme [2004]), lists (*e.g.*, Radev et al. [2000b]) or even complex structures such as hypertext (*e.g.*, White and Cardie [2002]).

Medium - Output medium is in most cases text although, as is the case with input media, output can consist of other media such as images and audio.

Other Output Factors - There can be a range of other output factors, such as derivation, genre, language, style. A discussion of each of these is provided by Sparck Jones [2007].

2.5.3 Approaches to Summarisation

Early Summarisation Approaches

To date there has been a considerable amount of research on automatic text summarisation using a variety of approaches. Providing a complete review of all summarisation systems and approaches

is not possible, and therefore, we provide an overview of the important advances made in this area in somewhat chronological order. We cover single- and multi-document summarisation and also extractive and non-extractive approaches. Almost all the approaches mentioned in this subsection are collectively covered by several surveys of this research area [Afantenos et al., 2005, Sparck Jones, 1999, 2007, Das and Martins, 2007, Zweigenbaum et al., 2007, Athenikos and Han, 2009].

The earliest known work on automatic summarisation dates back to 1958 when Luhn [1958] proposed that the *frequency* of words could provide a measure of their importance in a document. In his work, he ranked words based on their frequencies and used the ranking of individual words in a sentence to calculate their *significance*, finally selecting the top ranked sentences as the summary. Baxendale [1958] used *sentence position* as a feature for selecting important sentences in a document (*e.g.*, for news sentences, early sentences are more important than later sentences). Edmundson [1969] extended this work on summarisation and defined the framework for much of the work on extraction in what is known as the *Edmundsonian Paradigm* [Mani, 2001]. He used a linear function of four features to rank sentences for extraction:

$$W(s) = \alpha C(s) + \beta K(s) + \gamma L(s) + \delta T(s) \tag{2.1}$$

where α, β, γ and δ are manually assigned weights; W is the overall weight of sentence s, C represents the score given to sentence s due to the presence of cue words (bonus words or stigma words) extracted from a corpus, K represents the score given for key words (based on word frequency), L is the score given based on sentence location features, and T is the weight assigned based on terms in the sentence that are also present in the title. Such a statistical approach to sentence selection has been very popular in the research community with Earl [1970] being the first to investigate more varied and elaborate lexically based sentence features. More recently, numerous statistical approaches using multiple words, noun phrases, main verbs, and named entities. have been proposed [Barzilay and Elhadad, 1997, Lin and Hovy, 2000, Harabagiu and Lacatusu, 2005, Hovy and Lin, 1999, Lacatusu et al., 2003, Filatova and Hatzivassiloglou, 2004], while some research focused on utilising lexical items with 'importance-signalling' properties (*e.g.*, subheadings) [Teufel and Moens, 1997, Chakrabarti et al., 2001] or via the analysis of topical content [Ando et al., 2000]. The importance of the discourse structure of a text was realised quite early in summarisation research [Hearst, 1994, Marcu, 1997, 1998, Hahn and Strube, 1997] and this property has been heavily exploited ever since.

Recent Advances in Automatic Summarisation

Widespread research in automatic summarisation began from the mid 90's with the Dagstuhl Seminar [Endress-Niggemeyer et al., 1995], the Association for Computational Linguistics

Workshop in 1997, and primarily the Document Understanding Conference (DUC)[26] program. DUC was the first sustained evaluation program for automatic summarisation, and it was based on a broad road map which has been revised several times (further details of the roadmap is provided by Sparck Jones [2007]). Following the early summarisation research works, research on summary generation and evaluation techniques have been boosted by the DUC and similar other regular workshops and initiatives. They include: the NII Test Collection for Information Retrieval Project (NTCIR)[27], the Text REtrieval Conference (TREC)[28], and the Text Analysis Conference (TAC)[29], which was initiated from the Text Summarisation track of the DUC and the Question Answering track of the TREC. Funded by the National Institute of Standards and Technology (NIST) and other government agencies of the U.S.A., the TAC aims to support research within the NLP community by providing the infrastructure necessary for large-scale evaluation of NLP methodologies.

Since the commencement of widespread research on automatic text summarisation, a significant amount of research has focused on the application of machine learning algorithms for text summarisation. In most of the initial research, machine learning played a preliminary and support role — to identify information to be applied at specific process stages [Sparck Jones, 2007]. Early research mostly assumed feature independence and used the *Naïve Bayes* classifier [Kupiec et al., 1995, Lin and Hovy, 1997, Aone, 1999] with various features including those used in the *Edmundsonian Paradigm*. However, later research introduced the use of richer feature sets and a range of machine learning algorithms. For example, Lin [1999] used *Decision Trees*, Conroy and O'Leary [2001] used *Hidden Markov Models*, Osborne [2002] applied a *Log Linear Model* for sentence extraction, Svore et al. [2007] used *Neural Nets*, and Schilder and Kondadadi [2008] used *Support Vector Machines*. Interested readers can find more information about these approaches and other important related work in Sparck Jones [2007] and Das and Martins [2007].

While one branch of research focused on machine learning approaches, another branch progressed research on natural language analysis techniques. Miike et al. [1994] and Marcu [1998, 1999, 2000] use the Rhetorical Structure Theory (RST)[30] by building RST source text trees via the exploitation of discourse marker expressions in particular. They then use these trees to identify nucleus source clauses to extract for summaries. Polanyi et al. [2004] and Thione et al. [2004], in contrast, use the PALSUMM model which uses more abstract discourse trees. Barzilay and Elhadad [1997] show the use of *lexical chains* — sequences of related words that can span short or long distances — for single-document summarisation. The use of application-oriented structures has been more effective than the use of semantic discourse structures, especially within

[26] http://duc.nist.gov/. Accessed on 26th May, 2014.
[27] http://research.nii.ac.jp/ntcir/index-en.html. Accessed on 26th May, 2014.
[28] http://trec.nist.gov/. Accessed on 26th May, 2014.
[29] http://www.nist.gov/tac/. Accessed on 26th May, 2014.
[30] See http://www.di.uniba.it/intint/people/fior_file/INTINT05/RST.pdf. Accessed on 26th May, 2014.

particular domains [Sparck Jones, 2007]. For example, McKeown et al. [1998] and Elhadad and McKeown [2001] use a template suited to the medical domain; McKeown and Radev [1995] and Radev and Mckeown [1998] fill up template slots from a database of news information as a first step to their algorithm for news summarisation; and Sauper and Barzilay [2009], quite recently, show the use of templates to automatically generate Wikipedia articles.

Multi-document Summarisation

The field of multi-document summarisation was pioneered by the NLP community at Columbia University with the SUMMONS [McKeown and Radev, 1995, Radev and Mckeown, 1998] summarisation system developed in 1995. SUMMONS extended the already existing technology for template-driven message understanding systems. Extractive summarisation systems have been shown to work well for multiple documents particularly in the news domain, an example being the MEAD[31] [Radev et al., 2000a] system. As mentioned earlier, multi-document summarisation approaches suffer from the problems of incoherence and redundancy, and a number of approaches have been proposed in the literature to address these problems. One popular approach to reduce redundancy is the use of *clustering*. In this technique, common themes or concepts across document sets are identified and grouped or clustered together. Once the clusters are created, the summary can be generated by applying various algorithms that depend primarily on the content and compression needs. For example, some use a single sentence to represent each cluster in the final summary [McKeown and Radev, 1995, Radev et al., 2000b, Yih et al., 2007], while some generate a composite sentence from each cluster through the use of information fusion in order to combine the most important concepts from multiple sentences within a cluster [Barzilay and Elhadad, 1997, Barzilay et al., 1999, Barzilay and McKeown, 2005].

The Maximal Marginal Relevance (MMR) measure [Carbonell and Goldstein, 1998] is also commonly applied to reduce redundancy, particularly for query-driven summarisation. In this technique, relevant sentences are rewarded and redundant ones are penalised simultaneously by considering a linear combination of two similarity measures. The technique, thus, produces a set of relatively non-redundant but relevant sentences in the final summary. MMR was initially utilised for document retrieval and is given by the following formula:

$$MMR \equiv argmax_{D_i \in R \setminus S}[\lambda(Sim_1(D_i, Q) - (1-\lambda) max_{D_j \in S} Sim_2(D_i, D_j))] \tag{2.2}$$

where Q is a query; $R = IR(C, Q, \theta)$, i.e., the ranked list of documents retrieved by an IR system given a document collection C and a relevance threshold θ; S is the subset of documents in R already selected; $R \setminus S$ is the set difference, i.e., the set of unselected documents in R; Sim_1 is the similarity metric used in document retrieval and relevance ranking between documents and a

[31] Available at: http://www.summarisation.com/mead/. Accessed on 26th May, 2014.

query; and Sim_2 can be the same or a different metric.

Graph based approaches have also been applied to text summarisation [Mani and Bloedorn, 1997, Mani and Maybury, 1999, Erkan and Radev, 2004, Leskovec et al., 2005], with Mani and Bloedorn [1997] being the pioneers in this area. In their approach, the authors use nodes to represent words and edges between nodes represent relationships. The summaries generated can be topic driven, and no textual summary is generated. Instead, the summary content is represented via entities and relations (nodes and edges). The topic nodes are identified after a graph is created, and a search for semantically related text is propagated from the topic nodes to other nodes of the graph, in a process called *spreading activation*. When summarising a pair of documents, *common nodes* represent same words or synonyms, while *difference nodes* are those that are not common. Sentence selection from the graph is based on the average activated weights of the covered words: for a sentence s, its score in terms of coverage of common nodes is given by the following formula:

$$score(s) = \frac{1}{|c(s)|} \sum_{i=1}^{|c(s)|} weight(w_i) \tag{2.3}$$

where $c(s) = \{w|w \in Common \bigcap s\}$. The score for *differences* is calculated similarly. The sentences with higher *common* and *difference* scores are selected for the final summary. There have been some graph based approaches to summarisation in the medical domain, and they are discussed in the next subsection.

Erkan and Radev [2004] presented the LexRank system for multi-document summarisation, which uses a fully connected and undirected graph for the set of documents to be summarised. A similar method, suitable for single-document summarisation only, was proposed by Mihalcea and Tarau [2004]. Other graph based approaches have been proposed, both in the medical domain [Shi et al., 2007, Reeve et al., 2007, Fiszman et al., 2004] and outside it [Litvak and Last, 2008].

To end this brief review of text summarisation techniques, it is worth mentioning the centroid-based summarisation technique proposed by Radev et al. [2000a, 2004]. This technique is used in the MEAD system that has already been mentioned, and unlike a number of other systems, it does not make use of a language generation module. The summarisation is done in three stages. The first stage involves topic detection, with the goal to group together news articles that describe the same event. The centroid for each group or cluster is computed, where a centroid can be regarded as a pseudo-document that includes those words whose Term Frequency-Inverse Document Frequency ($tf \times idf$) scores are above a threshold in the documents that constitute the cluster. The two terms are given by the following equations:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{kj}}; idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{2.4}$$

where $n_{i,j}$ is the number of occurrences of the considered term t_i in document d_j, and the denominator is the sum of the number of occurrences of all terms in document d_j; |D| is the total number of documents in the corpus, and $|\{d : t_i \in d\}|$ is the number of documents where the term t_i appears.

The second stage uses the centroids to identify sentences in each cluster that are central to the topic of the entire cluster. The final score for each sentence is a combination of three scores (*centroid value*, *positional value* and *first-sentence overlap*) minus a redundancy penalty for each sentence that overlaps highly ranked sentences.

We now focus on the much narrower yet quite broad domain of medical text summarisation.

2.6 Summarisation and Question Answering for the Medical Domain

In this section, we provide an overview of text summarisation and Question Answering (QA) approaches targeted towards the medical domain. The goal here is to present some of the important research work in this area and analytically review more recent and promising approaches. The medical domain itself is quite broad, and so we attempt to adhere to research work that is relevant for evidence-based medicine. It must be mentioned that, in the recent past, there has been steady ongoing research in biomedical and medical text processing [Afantenos et al., 2005, Zweigenbaum et al., 2007]. However, compared to other domains, there is very little published research in summarisation and question answering [Fiszman et al., 2009]. We first provide a broad overview of approaches in the medical domain, and then provide a more elaborate analysis of specific systems that are most relevant to the research described in this thesis.

2.6.1 NLP Tools for Summarisation

There are a number of tools and resources capable of supporting NLP research in the medical domain. Here we briefly discuss some of the most important ones.

2.6.2 UMLS

The Unified Medical Language System (UMLS)[32] is a repository of biomedical vocabularies developed by the U.S. National Library of Medicine [Bodenreider, 2004]. The purpose of UMLS is to facilitate the development of computer systems that behave as if they *understand* the meaning

[32]http://www.nlm.nih.gov/index.html. Accessed on 26th May, 2014.

of the language of biomedicine and health. More specifically, it was developed as an effort to overcome two significant barriers to the effective retrieval of machine-readable information [Lindberg et al., 1993]: the variety of names used to express the same concept and the absence of a standard format for distributing terminologies. The UMLS consists of the following three major components:

- Metathesaurus – This is the major component of the UMLS and consists of a repository of inter-related biomedical concepts and terms from several controlled vocabularies and their relationships. It contains over a million biomedical concepts and five million concept names. It is the largest metathesaurus in the biomedical domain, providing a representation of biomedical knowledge consisting of concepts classified by semantic type and both hierarchical and non-hierarchical relationships among the concepts [Aronson, 2001]. Some examples of the controlled vocabularies are ICD-10, MeSH, SNOMED CT, LOINC and so on.

- Semantic Network – This provides a set of high-level categories and relationships used to categorise and relate the entries in the Metathesaurus. Each concept in the Metathesaurus is assigned to at least one 'semantic type' and certain 'semantic relationships' may be present between members of the various semantic types. For example, a *disease* mention (*e.g.*, headache) may have an *is-treated-by* relationship with a *drug* mention (*e.g.,* aspirin).

- SPECIALIST Lexicon – This is a database of lexicographic information for use in NLP. Each entry in it contains syntactic, morphological and orthographic (spelling) information. For example, a query on 'anesthetic' would return two entries with the part of speech tags noun and adjective, and its spelling variants.

SNOMED CT

SNOMED CT[33] is the most comprehensive source of medical terminology and is a U.S. standard for electronic health information exchange. It is accessible through NLM and the National Cancer Institute (NCI). It is one of the controlled vocabularies used by the UMLS.

MetaMap

The MetaMap[34] program was developed at the U.S. National Library of Medicine to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts

[33] http://www.ihtsdo.org. Accessed on 26th May, 2014.
[34] http://metamap.nlm.nih.gov. Accessed on 14th May, 2014.

referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, NLP and computational linguistic techniques.

SemRep

SemRep[35] is a program developed at the U.S. National Library of Medicine. It extracts semantic predications (subject-relation-object triples) from biomedical text. It has been used for a variety of biomedical applications, including automatic summarisation, literature-based discovery and hypothesis generation.

Other Tools

There are other text processing tools that have been used in the past for medical text processing although their functionalities are not restricted to this domain. The following is a brief list:

- GATE[36]: It is a Java platform from the University of Sheffield, and, according to their website, it is capable of solving almost any text processing problem.
- LingPipe[37]: A useful, Java-based language processing tool that is free for non-commercial use and is widely used for research.
- OpenNLP[38]: It is a source for a variety of Java-based NLP tools that can perform a range of basic and advanced text processing tasks.

2.6.3 Overview of the Medical Domain and Decision Support Systems

A number of factors make the medical domain a complex and interesting one for text processing. They include: large volume of data (*e.g.*, about 22 million articles in Medline alone); highly complex domain-specific terminologies (*e.g.*, drug names and disease names); domain specific lexical, terminological and ontological resources (such as UMLS); software tools and methods for exploiting the semantic information available (such as MetaMap [Aronson, 2001]); and domain specific format and typology of questions [Athenikos and Han, 2009]. Text in this domain mostly consists of clinical notes (often narrative) which are created by healthcare providers with limited time and also published medical articles describing studies/experiments and their outcomes. Only the latter type is relevant for evidence-based medicine, since the former cannot be considered as a form of reliable medical literature and is also generally not publicly available.

[35] http://semrep.nlm.nih.gov. Accessed on 14th May, 2014.
[36] http://gate.ac.uk. Accessed on 17th May, 2014.
[37] http://alias-i.com/lingpipe. Accessed on 18th May, 2014.
[38] http://opennlp.sourceforge.net. Accessed on 25th May, 2014.

The integration of technology with medical practice was initiated through the use of Clinical Decision Support (CDS) systems. Such systems help practitioners make clinical decisions, deal with medical data or with the knowledge of medicine necessary to interpret such data [Shortliffe, 1990]. CDS systems were introduced about four decades ago, and, according to studies, they have improved practitioner performance in approximately 60% of the one-hundred reviewed cases [Garg et al., 2005]. Early CDS systems consisted mostly of applications that facilitated diagnosis, treatment [Shortliffe et al., 1979, Miller et al., 1982] and the management of patients (*e.g.*, through computerised guidelines and alerts) [McDonald, 1976, Barnett et al., 1983]. However, the capabilities of such applications were quite limited, primarily because they did not have access to raw medical data [Friedman and Hripcsak, 1999].

Since a lot of medical data is textual, the need to integrate NLP with CDS systems became more noticeable as the volume of medical data increased. However, this requirement introduced new challenges as well, since CDS systems that rely on NLP require reliable, high quality NLP performance and modular, flexible and fast systems, as Demner-Fushman et al. [2009] points out. Research in this area made rapid progress since the mid 90's, and today there is a wide range of NLP-based CDS systems in use (some of these are discussed later in this chapter). A detailed discussion of CDS systems is outside the scope of this chapter and in the following subsections, we focus instead on a subset of CDS systems that attempt to provide answers to clinical queries by summarising the information contained in medical texts. Note that such systems consist of information extraction, summarisation, and QA components, and we do not distinguish between these three types. Identifying and presenting evidence in a condensed manner is essentially a task of summarisation. Hence we refer to these systems/components as summarisation systems. For QA systems, we primarily discuss their summarisation components. For the interested reader, Friedman and Hripcsak [1999] provide a review of early NLP-based CDS approaches and a recent and detailed analysis of CDS systems and NLP for the medical domain is provided by Demner-Fushman et al. [2009]. Readers unfamiliar with QA may refer to Mollá and Vicedo [2007] for an overview of QA techniques in restricted domains and Athenikos and Han [2009] for a review of QA approaches in the biomedical domain.

2.6.4 Overview of Medical Text Summarisation Approaches

Information Extraction Approaches

Early summarisation systems were mostly concerned with extracting relevant information from structured or unstructured medical text. The Linguistic String Project (LSP) [Sager et al., 1994] is an example of early medical NLP work. Its primary purpose is the transformation of clinical narratives into formal representations. The system named 'Medical Language Processor' (MLP), is not a pure summarisation system, although it does apply NLP techniques for information ex-

traction and content selection. The system performs five stages of processing – parsing, selection, transformation, regularisation and information formatting. MedLEE [Friedman, 2005] is also responsible for extracting information from clinical narratives and presenting the information in structured forms through the use of a controlled vocabulary. It is used for processing various forms of notes and reports and is integrated with a clinical information system. HiTEx [Zeng et al., 2006] is yet another, relatively recent, system used for extracting findings such as diagnoses and family history from clinical narratives through the use of NLP techniques. TRESTLE (Text Retrieval Extraction and summarisation Technologies for large Enterprises) [Gaizauskas et al., 2001] is essentially an information extraction system that produces single sentence summaries of pharmaceutical newsletters through the identification of Named Entities (NEs) followed by sentence extraction based on the presence of key NEs. Drug and disease names are considered by this system to be named entities. Hahn et al. [2002] present MEDSYNDIKATE, a natural language processor that automatically acquires medical information from findings reports. The contents of the texts are transferred to conceptual representations that correspond to a knowledge base, and the system incorporates domain knowledge to semantically interpret major syntactic patterns in medical documents. Xu et al. [2010] propose the MedEx system which uses NLP to extract medication information from clinical notes with very high recall and precision. Numerous other information extraction techniques have been proposed for medical texts, with various intents such as: named entity recognition, adverse drug event detection, negated concept detection, and many more.

Extractive Summarisation Approaches

Most summarisation systems in this domain applied extractive summarisation approaches like the initial MiTAP system (MITRE Text and Audio Processing) [Damianos et al., 2002] which was targeted towards the monitoring of infectious disease outbreaks or other biological threats. MiTAP monitors various sources of information such as online news, television news, and newswire feeds. and captures information which are filtered and processed to identify sentences, paragraphs, and POS tags. The final summary is generated by WebSumm [Mani and Maybury, 1999] as extracted sentences from the processed text. Johnson et al. [2002] also present an extractive summarisation approach in which medical documents are clustered into groups, which are then analysed for features and a cluster signature is generated. The summary is generated by matching the cluster signature to each sentence of the document to be summarised, and ranked sentences are presented to the user as a summary. Reeve et al. [2007] propose a single-document, extractive summarisation approach that combines BioChain [Reeve et al., 2006a], which identifies relevant sentences using concept-chaining (similar to lexical chaining but applied to UMLS concepts), and the FreqDist system [Reeve et al., 2006b], which uses a frequency distribution model to identify relevant sentences. This hybrid approach, called ChainFreq, first uses the BioChain method

to identify candidate sentences containing strong concepts. The candidate sentences and their corresponding concepts are then passed to the FreqDist method, which produces a set of summary sentences from the candidate sentences. In the second step, sentences are selected in a way such that the frequency distribution of the concepts in the summary is similar to that in the original text. Other extractive approaches have been proposed with various intents: Mihalcea [2004] presents an approach for automatic sentence extraction using graph based ranking algorithms; Elhadad [2006] proposes extractive algorithms for performing user-sensitive text summarisation; and, more recently, Ben Abacha and Zweigenbaum [2011] put forth some summarisation approaches for the automatic extraction of semantic relations between medical entities.

Non-extractive Approaches

MUSI is an early system that applies semantic information thoroughly [Lenci et al., 2002] (MUltilingual summarisation for the Internet). It follows an approach similar to the ones already mentioned for sentence extraction but also has the capability of producing semantic representations of the extracted sentences to produce an abstractive summary. In the case of an abstractive summary, a range of text processing techniques including tokenisation, morphological analysis, shallow syntactic parsing, chunking, dependency analysis and mapping to the internal representation is performed. The final summary is produced in French and German with the aid of Natural Language Generation systems.

Other than the MUSI system, early research in this area did not consider abstractive summarisation. Importantly, most of the mentioned systems are applicable to single documents, and progress in multi-document summarisation in this domain has been made over the last few years only. Recently, a number of abstractive summarisation systems have been proposed (*e.g.*, PERSIVAL [Elhadad and McKeown, 2001], MedQA [Lee et al., 2006a, Yu et al., 2007a], EpoCare [Niu et al., 2005, 2006]. We discuss some of these approaches in more detail in the following subsections.

Progress in Medical Summarisation and Question Answering

In recent years, medical document summarisation has progressed in a variety of directions. While some systems rely on heavy linguistic processing, most make use of domain knowledge and/or statistical techniques. A variety of document types are now addressed by the different approaches, and QA systems have been developed specific to this domain.

Clinical notes, narratives and recorded patient information have remained important summarisation subjects. Meng et al. [2005] use extractive summarisation to obtain patient information using semantic patterns, and Vleck et al. [2007] identify information physicians consider relevant to summarising a patient's medical history in a medical record. More recently, Sarkar

[2009] presents an approach to summarise published medical documents taking advantage of domain-specific information provided by UMLS. The proposed approach is extractive, based on the *Edmundsonian Paradigm* and uses features such as term frequency, sentence similarity to document title, position of sentence, presence of domain specific cue phrases, presence of novel terms (terms not found in either domain-specific or domain-independent vocabulary sources) and sentence length.

Most work on query-focused summarisation in this domain has been carried out under the broader research area of QA. Initial QA research has focused on open domain text-based QA. The field has witnessed a growing interest among researchers in restricted domain QA [Mollá and Vicedo, 2007]. In contrast to open domain QA, QA in the biomedical and medical domains are challenged with a more acute need to cater for specialised terminological variation [Zweigenbaum, 2003]. Zweigenbaum [2009] also notes the role of domain-specific knowledge and reasoning for restricted domain (such as the medical domain) QA. The author states that knowledge and reasoning are both more necessary and manageable for a restricted domain compared to an unrestricted domain due to the relative specificity, and the difficulty of limited scoping of questions.

Some research has focused primarily on question analysis and classification as a first step towards medical QA [Athenikos and Han, 2009]. Yu and Sable [2005], Yu et al. [2005, 2007a] and Yu and Cao [2008], for example, study the answerability of clinical questions and attempt to classify clinical questions based on the Evidence Taxonomy [Ely et al., 1999], and also into general topics. Terol et al. [2007] present a very similar approach and propose a logic-form based approach in adapting a generic restricted-domain QA system to the medical domain. The primary component of this system is the question analysis component that classifies each question to one of the 10 most generic questions proposed by Ely et al. [2000]. From the question, logic forms representing the information needs of the question and the answer pattern are derived. The answer extraction phase, then becomes the problem of finding the relevant concepts from the documents that can fulfill the needs of the question. The evaluation of the system, however, only evaluates the question analysis phase and therefore the performance of the overall system remains.

The QA approach proposed by Weiming et al. [2007] is quite different and relies heavily on semantic information present on the questions and documents. In the answer generation phase, candidate sentences are first identified from which phrase-level answers are generated by mapping semantic types and relations in the candidate sentences to those in the question. For each answer generated by semantic clustering, the system provides the semantic type, associated concepts and the sentence from which the answer originates. The system has been evaluated for factoid and complex questions and is shown to have very good recall and precision (77% and 92% respectively). Finally, Workman et al. [2012] present a dynamic summarisation approach using

an algorithm called *Combo* to identify salient semantic predications. It was shown to outperform several baseline methods in terms of recall and precision.

2.6.5 Detailed Review of Systems: The Evidence-based Medicine Perspective

While discussing approaches in the previous subsection, we attempted to only provide the reader with a flavour of the various directions in which medical text summarisation has progressed. We intentionally skipped some important systems and filtered out many. In this subsection, we discuss in detail the characteristics of a number of systems that perform summarisation of medical text. For full QA systems, we primarily focus on their summarisation components. This review not only discusses the capabilities of the systems but also presents their strengths and weaknesses from the perspective of evidence-based medicine. Table 4.3 provides a summary comparison of the systems mentioned here.[39]

[39]Note that in the table, SemRep represents the summarisation approach using SemRep proposed by Fiszman et al. [2009].

System	Input	Unit	Use of Semantic Information	Summary Type	Target user
MedQA*	Medline and the Web	Multi-document	No	Non-extractive/Definitional	Healthcare Practitioners
CQA 1.0*	Medline abstracts	Multi-document but separate summaries for each abstract	Yes (UMLS)	Extractive	Healthcare Practitioners
BioSquash	Medline abstracts	Multi-document	Yes (UMLS and WordNet)	Extractive	Unspecified
SemRep	Clinical trials from Medline	Multi-document	Yes (UMLS and SemRep)	Non-extractive	Healthcare Practitioners
EpoCare	Medline abstracts (cited by CE articles)	Multi-document	Yes (UMLS)	Extractive	Healthcare Practitioners
PERSIVAL	Patient Records, Medical Articles and Web-based Text Articles	Multi-document	Yes (UMLS)	Non-extractive	Healthcare Practitioner and Laymen
AskHermes*	Medline abstracts and full texts	Multi-document	Yes (UMLS)	Extractive	Healthcare Practitioners

Table 2.1: Comparison of summarisation systems for the medical domain. Systems available online are marked with a *.

MedQA

MedQA[40] [Lee et al., 2006a, Yu et al., 2007a] answers definitional questions by producing paragraph-level answers from Medline and the web. The system uses question classification [Yu and Sable, 2005, Yu et al., 2005] in an initial stage followed by a shallow syntactic parser and a standard IR engine for query formulation and document retrieval. Multiple strategies are used for answer extraction including document zone detection, sentence categorisation using cue phrases and identification of lexico-syntactic patterns that comprise definitional sentences. Hierarchical clustering [Lee et al., 2006b] and centroid-based summarisation techniques (already discussed) [Radev et al., 2000b] are used for text summarisation. Yu et al. [2007a] and Yu and Kaufman [2007] provide an evaluation of MedQA against three search engines – Google, OneLook and PubMed – for answering definitional questions and show that Google is very effective in obtaining definitions, outperforming MedQA.

MedQA has opened new directions in medical text summarisation. It has shown how supervised classification can be used for intermediate steps in medical text summarisation, such as query analysis. MedQA, however, does not capture semantic information which plays an important role for answer extraction and summarisation particularly for this domain. Also the intent of the system is very different from the requirements of the evidence-based medicine practitioner. Due to its limited capability and its narrow purpose, its role in evidence-based medicine practice is also limited. The key limitation of the system is that it can only answer definitional questions, while the questions that appear in real life evidence-based medicine practice are generally complex in nature. Despite its limitations, the techniques employed by the system are very useful for implementing QA systems in this domain.

CQA 1.0

Lin and Demner-Fushman [2007] present a QA system[41] that uses a statistical and knowledge-based approach and is particularly targeted towards the practice of evidence-based medicine. The proposed system uses PICO representations of questions as queries which are sent to PubMed to retrieve an initial set of abstracts. From the abstracts, each of the PICO elements (Problem/Population, Intervention, Comparison and Outcome) are extracted using various techniques. MetaMap is extensively used to identify UMLS terms and their categories. The population extractor uses a series of hand-crafted rules to identify occurrences of population terms in the abstracts, with preference given to terms occurring earlier in the documents. The problem extractor relies on the recognition of concepts belonging to the UMLS semantic group 'DISORDER'. It returns a list of problems, with the concepts appearing in the title, introduction or first two lines of the sentences

[40] Available at: http://askhermes.org/MedQA/. Accessed on 26th May, 2014.
[41] Available at: http://www.umiacs.umd.edu/~demner/. Accessed on 26th May, 2014.

given higher preference, and the highest ranked problem is chosen as the primary problem. The intervention and comparison elements are identified in a similar way – by recognition of nine semantic types (*e.g.*, *diagnostic procedure*, *clinical drug*, etc.). In structured abstracts, more weight is given if the semantic types appear in 'title', 'aims' or 'methods' sections while in unstructured abstracts, more weight is given if they appear towards the beginning of the document. Using the extracted knowledge, the authors re-rank the retrieved documents using a document ranking algorithm that takes into account the knowledge elements, strength of evidence and other task specific considerations [Lin and Demner-Fushman, 2006]. An outcome extractor extracts outcome sentences from the retrieved documents using a strategy based on an ensemble of classifiers (a rule-based classifier, a bag-of-words classifier, an n-gram classifier, a position classifier, an abstract length classifier and a semantic classifier). Each sentence is given a probability based on the classifier scores, and the top-ranked sentences are chosen. As the final output, the system simply produces the top ranked sentences from the top re-ranked documents along with the question and the strength of recommendation. Only basic clinical questions such as 'What is the best drug treatment for X' are addressed. The authors evaluate the performance of the knowledge extractors against different baselines and also manually evaluate the final output against a baseline that only presents top sentences from unranked documents.

The CQA 1.0 system is invoked as a module by the InfoBot system that is under development at the NLM [Seckman et al., 2008]. The techniques applied by the CQA system show the importance of statistics and domain-specific knowledge for medical text summarisation. There are several limitations of the system. It relies on PICO frames for queries (instead of natural language questions), and an information synthesis technique is absent at the end to produce a single answer from related documents. Furthermore, the algorithm applied to predict the qualities of evidences does not follow an evidence-based guideline. In our work, we utilise some of the ideas proposed by the authors of CQA 1.0, and we address some of its limitations.

BIOSQUASH

BioSquash is a question-oriented extractive summarisation system for biomedical documents [Shi et al., 2007, Melli et al., 2005]. The system has four main components – Annotator Module, Concept Similarity Module, Extractor Module and Editor Module. The Annotator Module uses a statistical parser, a named-entity recogniser and a semantic role labeler to annotate the documents with syntactic and shallow semantic information. The Concept Similarity Module obtains semantic meanings of biomedical concepts and ontological relations among these concepts. WordNet[42] and UMLS are used to extract concepts. The extractor module performs content selection in three steps – Content Identification, Text Graph Creation, Concept Significance

[42] http://wordnet.princeton.edu/. Accessed on 26th May, 2014.

Identification and Concept Space Covering. The Concept Identification step identifies ontological concepts, named entities and noun phrases. The Text Graph Creation step then represents relationships between concepts, even those in separate documents, as edges in a graph whose nodes are the concepts. The significance of each concept is then determined using the number of edges it is connected to and also based on the type of concept. Highest ranked propositions (groups of concepts) are then chosen as candidate sentences, and a penalty function is used to prevent similar sentences from being selected. The Editor Module reorders candidate sentences using a 2-phase re-ordering algorithm, and a summary candidate is generated after compressing the reordered sentences.

The approach presented is quite comprehensive and makes heavy use of domain knowledge and statistical techniques. The system is evaluated on only 18 questions from TREC-2005[43] using ROUGE[44] and is shown to perform reasonably well. However, such a small test set is not sufficient to evaluate the full capabilities of the system, and a more comprehensive evaluation is desirable.

Summarisation using SemRep

Fiszman et al. [2004], Rindflesch et al. [2005] and Fiszman et al. [2009] propose a semantic abstraction approach to automatic summarisation in the biomedical domain by using SemRep as a semantic processor to perform source interpretation and predication listing. Summarisation relies on a user-specified topic. A transformation stage generalises and condenses the list of predicates generating a conceptual condensate for the input topic. The transformation is carried out in four phases: Relevance (include predications in the topic of summary), Connectivity (also include 'useful' additional predications), Novelty (do not include predication that the user already knows) and Saliency (only include the most frequently occurring predications). The Relevance phase identifies predications on a given topic (*e.g.*, disorders) and is controlled by a schema represented as a set of predications in which the predicate is drawn from a relation in the UMLS Semantic Network, and the arguments are represented as a domain covering a class of concepts in the Metathesaurus. Predications produced by SemRep must conform to this schema in order to be included in the conceptual condensate; such predications are called 'core predications'. The Connectivity phase is a generalisation process and retrieves all predications that share arguments with core predications. The Novelty phase provides further condensation by eliminating predications that have a generic argument (*e.g.*, Pharmaceutical Preparations) as determined by its hierarchical depth in the Metathesaurus. Finally, the Saliency phase calculates the occurrence of arguments, predicates and predications, and those occurring more frequently

[43] http://ir.ohsu.edu/genomics/. Accessed on 26th May, 2014.

[44] ROUGE — Recall-Oriented Understudy for Gisting Evaluation — is discussed in Section 7

than others in the condensate are kept while the others are eliminated. The final summary is produced in the form of a graph, and the approach can be applied to both single documents and multiple documents without requiring any modifications.

One of the drawbacks of the system is that it does not incorporate query information. The system is evaluated [Fiszman et al., 2009] on its ability to identify drug interventions for diseases only (other forms of interventions are not taken into account). Therefore, the application domain of the system is very limited. Furthermore, only clinical trials are used as source documents. The performance of the system is compared to a baseline that selects drug names based on the frequencies of their occurrence in source texts. A scoring mechanism called the *clinical usefulness score* is used. It rewards the systems for identifying beneficial drug interventions and penalises them for identifying harmful or not useful ones. To assess the usefulness of drug interventions, drug listings in *Clinical Evidence* (CE) articles are used as surrogates for a physician-annotated reference standard. The system is shown to outperform the baseline in terms of both the clinical usefulness score and mean average precision (MAP).

The summarisation approach applied by the SemRep system is simple, innovative, and effective. Its performance illustrates the importance of domain-specific semantic information, and the usefulness of distributional semantics in automatic summarisation for this domain. Incorporating query-focus into the summarisation procedure will make the technique more applicable for evidence-based medicine practice. Summary information requirements must be identified from clinical questions instead of manually identified topics, and the summarisation component should be able to identify information other than drug interventions.

EPoCare

The EpoCare (Evidence at Point of Care)[45] project [Niu et al., 2003, Niu and Hirst, 2004, Niu et al., 2005, 2006] is an initiative by the University of Toronto to develop a system that can automatically answer clinical questions. The project aims to provide fast access at point-of-care to the best available medical information. The current implementation relies heavily on automatic identification of PICO elements from both clinical questions and their corresponding answers. PICO keywords are first identified from the question and used as keywords for retrieval. The problem of identifying answers to a clinical question is divided into four sub-problems – (i) identifying roles (PICO elements) in the text, (ii) determining the lexical boundary of each role, (iii) analysing the relationship between different roles and (iv) determining which combinations of roles are most likely to contain correct answers. The initial work presented by Niu and colleagues addresses simple treatment-type questions. MetaMap is used for automatic identification of

[45] http://www.cs.toronto.edu/km/epocare/index.html. Accessed on 26th May, 2014.

interventions and problems. The authors note that identification of outcomes is a much more difficult task, and they identify cue words (nouns, verbs and adjectives) that indicate the presence of outcomes in sentences. The outcome detection task is further subdivided into two sub-tasks — outcome identification and lexical boundary determination. In addition to sentence level outcome detection, the authors suggest that the polarity of outcomes play an important role in determining which sentences to choose as answers. Four categories of polarities are defined (positive, negative, no outcome and neutral). The authors use a Support Vector Machines (SVM) to classify sentences into the four categories and use uni-grams, bi-grams, change phrases, negations and semantic categories (UMLS) as features. Upon assessment of the effect of different combinations of features, the authors show that best results are obtained by combining linguistic features with domain features.

The outcomes of the polarity classification task are used in a multi-document summarisation approach to automatically find information from Medline abstracts to answer a clinical question. Presence and polarity of outcomes, position of sentence in abstracts, length of sentences and Maximal Marginal Relevance (MMR) are used as features in a machine learning algorithm (SVM) used to solve the summarisation problem. Sentences in Medline abstracts cited by Clinical Evidence (CE) articles are manually annotated for each clinical query, and sentences obtained from the automatic approach are compared with these for evaluation. A total of 197 abstracts from 24 subsections in CE are annotated to give a total of 2,298 sentences.

The outcome detection task is shown to have an accuracy of 83%. The polarity assessment task is shown to have an accuracy of 79.42%. For the summarisation task, it is observed that the identification of outcomes and polarity improves performance significantly. However, F-scores are shown to be very low (<0.50) in all cases. ROUGE is also used for evaluation but shows little difference in performance of different combination of features.

The outcomes of the EpoCare project show that abstractive summarisation approaches, which utilise automatic polarity classification of sentences, have the potential to be applied for the generation of evidence-based summaries. This work is very promising, and future research on the incorporation of *context* information (*e.g.*, polarity relative to a context intervention, and contexts specified by queries) appear lucrative. Due to the promise of this approach, we build on the initial work of the EPoCare project for the generation of bottom-line, evidence-based recommendations. Our work related to this is described in Chapter 6.

PERSIVAL

PERSIVAL (PErsonalized Retrieval and summarisation of Image, Video and Language) [Elhadad and McKeown, 2001] is a medical digital library designed to provide personalised access to

a distributed library of multimedia medical literature. It is not possible to discuss the whole project in this review, and hence we focus on the text summarisation component of the system that produces customised, abstractive summaries for persons from technical and non-technical backgrounds [Elhadad et al., 2005, Elhadad, 2006].

The summarisation system takes as input three different sources: patient records, medical articles (about cardiology) that are relevant to the patient and the user query from which the key words are extracted. The input articles are classified as prognosis, treatment or diagnosis. Relevant sentences from the 'Results' sections of the articles are extracted and stored in tuples of the form: (Parameter(s), Findings, Relation), where the Relation describes the relation between a parameter and a finding. The information extraction phase also identifies the degree of dependence of the parameters, the article and the sentence from which it has been extracted and various other minor information [Elhadad et al., 2005]. Hand-crafted templates are filled with the extracted information, and, after identifying the portions of the extracted parameters that are relevant to the patient records, the templates are merged and ordered by rendering into an internal semantic representation in the form of a graph. Repetitions (when two nodes are connected via multiple vertices of the same type) and contradictions (when two nodes are connected by multiple vertices of different types) are identified from the graphs, and this information is used to create a coherent summary. The information is then ordered for summary generation: relations related to the user query are given the highest preference followed by recitations and contradictions, which in turn are followed by preference based on the relation type (*e.g.*, risk relation is given higher priority than association relation) and, finally, dependent relations from the same template are presented together. The final summary is generated with the use of NLG techniques, and all medical terms are hyperlinked to their definitions. The system is built through close collaborations with medical experts who check the validity and performance of the approaches at each iteration of development, and the authors use this informal evaluation at each stage to validate and improve their system. Each prototype of the system is built as an improvement to the previous one based on the feedback obtained from the experts.

The summaries generated by the system are not evidence-based, and the intent of the system is to provide personalised information for users from both technical and non-technical backgrounds. The approaches applied by this system have the potential of being applied for evidence-based summarisation.

AskHERMES

AskHERMES[46] [Cao et al., 2011] is a clinical QA system that performs robust semantic analysis of complex clinical questions and output query-focused extractive (single-document) summaries as answers. The system allows the users to enter questions using natural language with minimal query formulation, and to efficiently navigate among all the answer sentences to quickly meet their information needs. The system is demonstrated to outperform Google and UpToDate when answering complex clinical questions.

AskHermes operates in five phases: a *query analysis* module automatically extracts information needs from the questions and outputs a list of query terms; a *related questions extraction* module returns a list of similar questions; an *information retrieval* module returns the relevant documents; an *information extraction* module identifies relevant passages from the source documents; and a *summarisation and answer presentation* module aggregates answer passages, removes redundant information, and automatically generates structured summaries.

In the query analysis phase, a query is first classified into one of 12 general topics [Yu and Cao, 2008], and then *keywords* are automatically extracted from the question using Conditional Random Fields. Following the retrieval of relevant documents, a two-layer hierarchical clustering is applied to group passages into different topics. Topic labels are assigned to clusters using query terms and expanded terms from the UMLS. A topic-labeled tree structure is generated from the first layer of clustering, and a second layer of clustering is applied to provide more refined categories. Clusters are ranked based on the *key* query terms appearing in the cluster, and redundancies are detected and removed using *longest common substrings*.

AskHERMES is shown to perform comparably to state-of-the-art systems, and its potential in the field of evidence-based medicine is very promising. One issue is that the lengthy summaries do not satisfy the requirement of bottom-line recommendations required by practitioners. Due to the extractive nature of the summarisation, it is difficult to synthesise information from multiple documents and generate brief summaries. However, the system's performance suggests that customising the summary generation process to the type of question may be beneficial. Furthermore, the authors show that key query terms can be used to determine topics. This suggests that the semantic types of these key query terms, and the semantic associations they have with the document terms, may play an important role in query-focused summarisation.

[46] Available at: http://www.askhermes.org/index2.html. Accessed on 26th May, 2014.

2.7 Evaluation

The objective of this section is to briefly discuss the evaluation techniques used for automatic summarisation, approaches attempted in the past and possible approaches for evaluating summaries for evidence-based medicine. Evaluation of automatically generated summaries is a hard problem [Sparck Jones, 1999, 2007, Das and Martins, 2007]. This is primarily because of the fact that it is a heuristic problem – there are more than one acceptable solutions, and a universally accepted standard evaluation is absent. Evaluation measures usually focus on specific features of the summarised text, which depend largely on the summarisation factors. The features focused upon by evaluation techniques for generic summaries can be significantly different from those for query-focused summaries. Evaluation techniques can also vary according to the unit of summarisation (i.e., single *vs.* multi-document), domain, type (extractive *vs.* non-extractive) and other factors mentioned previously. In this brief section, we focus only on approaches that have been used or may be relevant for application in evidence-based medicine summarisers. The interested reader can refer to Lin and Hovy [2002] for a brief discussion on automatic and manual evaluation techniques.

2.7.1 Extrinsic and Intrinsic Evaluation

Summary evaluation techniques can be divided into two broad categories – extrinsic and intrinsic. Extrinsic evaluation techniques focus on the purpose of summaries. The objective is to measure the usefulness of a summary for a specific task. The single feature that plays the most important role in determining the usefulness of a summary is its content, and numerous evaluation techniques use this feature as an indication of the qualities of summaries. Recall, precision, F-score, coverage and other similar measures [Lin and Hovy, 2002] have been frequently used in the past as simple evaluations of content relevance (*e.g.*, Weiming et al. [2007], Niu et al. [2003], Niu and Hirst [2004], Niu et al. [2005, 2006]). More sophisticated techniques have also been applied for extrinsic evaluation. An example is the assessment of the answerability of questions (that can be answered from the source text) from summarised text: Morris et al. [1992] used educational reading comprehension questions; Minel et al. [1997] and Teufel [2001] use more sophisticated questions about argument structure; SUMMAC [Mani et al., 2002] used questions about significant source content that should be answerable from summaries. Such an evaluation approach can be useful for summarisers targeted towards evidence-based medicine, as discussed later.

Intrinsic methods concentrate on the summary itself, trying to measure features such as coherence, cohesion, grammaticality, readability and other important features. Intrinsic methods for extractive summaries assess features such as discourse well-formedness, while those for non-extractive

approaches must also assess sentence well-formedness. Although in some domains, such as news, intrinsic evaluation plays an important role, for a query-focused summarisation system for evidence-based medicine, we believe it to be much less important. The reasons are discussed later in the section.

2.7.2 Evaluation Techniques

Both manual and automatic evaluation techniques are used for evaluating summaries. The following are some common techniques used in both methods of evaluation.

Gold Standards

Gold standards (also known as human reference summaries) are often used for evaluating automatic summarisation primarily because humans can be relied upon to capture important source content and to produce well-formed output text [Sparck Jones, 2007, Suominen et al., 2008]. The expected output summaries are manually created by human experts, who are often experts in a specific domain. The created summaries therefore contain the necessary content and become the target performance for systems. Evaluation then compares the generated summaries with the gold standard summaries. This can be done automatically or manually, and is itself a research problem. The more similar a generated summary is to the gold standard, the better it is considered to be. In evaluation of early extractive summarisation approaches, the automatic summaries were compared against extracted gold sentences picked by human experts. Since then, this technique has expanded to evaluating domain-specific summarisation systems where gold summaries are generated by domain experts, and often customised comparison techniques are used. Automatic comparisons against gold summaries are also used, such as ROUGE [Lin and Hovy, 2003], which has been used to evaluate a number of the summarisation systems mentioned in this paper. Further details about this evaluation system are provided later.

For years, the absence of gold standard data has been an obstacle to summarisation research in the medical domain. Due to this reason, there has been very little research on analysing the contents of human summaries and data-driven summarisation approaches. Absence of gold standards has also made automatic, relative evaluation of systems difficult.

Baselines

Baselines can be considered to be the opposite of gold standards in that they indicate the minimum level of required performance by a summarisation system. For extractive summarisation, various baselines such as n random sentences have been used. A more appropriate baseline for news

summarisation was introduced by Brandow et al. [1995] who used the first *n* sentences. With these baselines, the minimum performance required by a system is to select *n* sentences that better summarise the source than the baseline. Similar baselines have been established for summarisation in various other domains. For example, Lin and Demner-Fushman [2007] propose an outcome extractor that is compared against a baseline of 'last *n* sentences' (since outcomes presented in a medical paper usually appear towards the end of the abstract). Baseline measures of tf.idf type have also been used in the literature and even for evidence-based medicine [Fiszman et al., 2009]. In such baselines, the summarisation units (sentences, words, n-grams) are chosen based on their tf.idf values. A standard baseline for summarisation systems across domains and even within specific domains, however, still does not exist.

Topic-oriented Evaluation

Topic-oriented evaluation techniques are specialised to the topic and intent of the summarisation task. A number of topic-oriented evaluation schemes have been proposed in the literature, both within the medical domain and outside. A recent example of a topic-oriented evaluation mechanism is the 'Clinical Usefulness Score' (CUS) [Fiszman et al., 2009], a unique evaluation of generated summaries. The CUS is a categorical performance metric. In calculating this score, interventions extracted by a system are assigned to one of four high-level categories depending on how they match the interventions in a predetermined reference standard. The goal is to give credit to a system for finding beneficial interventions and, similarly, penalise it for finding harmful interventions. The high-level categories and the corresponding reference standard categories are — best (beneficial), ok (trade-off between benefits and harms), bad (harmful) and other (unknown). Overall CUS for a topic is calculated as follows:

$$CUS = w_{best} \times bs_1 + w_{ok} \times os_1 - w_{other} \times os_2 - w_{bad} \times bs_2 \tag{2.5}$$

where bs_1 is the BEST score, bs_2 is the BAD score, os_1 is the OK score and os_2 is the OTHER score. The scores are simply the number of drugs identified that fall into each of the four categories divided by the total number of drugs identified for the whole topic. The weights are manually assigned.

2.7.3 Manual Evaluation

Due to the difficulty associated with evaluating automatic summaries, manual evaluation is still a common practice. There is more confidence in manual evaluation (compared to automatic

evaluation) since humans can infer, paraphrase and use world knowledge to relate to text units with similar meanings but worded differently. In such evaluations, domain experts (often several for a single summary) read and grade summaries, usually using some chosen scale. Manual pairwise comparison of generated summaries with gold standard summaries is also frequently used. Most of the systems mentioned in the last subsection of the previous section have undergone some form of manual evaluation, or at least involved human experts for the preparation of a gold standard (*e.g.*, [Lin and Demner-Fushman, 2007]). However, agreement among human summarisers is generally quite low, and the process of manual evaluation is quite expensive. Human judgements have been shown to be unstable and inconsistent as well [Lin and Hovy, 2002]. As a result, alternative automatic evaluations having high correlation with human scores are usually used.

Nenkova and Passonneau [2004] present the pyramid approach for manual summary evaluation. According to the authors, instead of attempting to elicit reliable judgement from humans, this evaluation method is calibrated to human summarisation behaviour. Summary content is divided into summarisation content units (SCU), and SCUs representing the same semantic information are annotated in each source document. Once annotation is complete, each SCU is given a weight equal to the number of summaries in which the SCU appears. Next, the SCUs are partitioned into a pyramid in which each tier contains SCUs of the same weight and higher tiers contain SCUs of higher score (i.e., SCUs appearing in more human summaries). Therefore, an optimal summary is expected to contain SCUs from the top tier followed by (if length permits) SCUs from the next tier and so on. Finally, the score assigned to an automatically generated summary is the ratio of the sum of the weights of its SCUs to the sum of the weights of an optimal summary with the same number of SCUs.

A slightly earlier approach presented by van Halteren and Teufel [2003] is similar: in it atomic semantic units, called factoids, are used to represent the meanings of sentences. The approach requires the generation of a large number of summarised articles from which the gold standard can be obtained by identifying the most frequently occurring factoids. As an example, the authors show that to generate a 100 word summary of a news article (from 50 sample summaries), all factoids appearing in at least 30% of the summaries had to be included. Hence, gold standards of different lengths can be generated by varying the factoid threshold. Although this approach ensures very high agreement among the manual summarisers, its requirement of a large number of sample summaries makes it quite an expensive approach.

2.7.4 Automatic Evaluation

ROUGE

Lin and colleagues [Lin and Hovy, 2003, Lin, 2004] introduced a set of metrics called Recall-Oriented Understudy for Gisting Evaluation (ROUGE) that have become very much the standards of automatic evaluation. The intent of the various ROUGE measures is to find the similarity between automatically generated summaries and reference summaries. One of the metrics, ROUGE-N, is an n-gram based recall oriented statistic and can be used with multiple reference summaries. The statistic is calculated using the following formula:

$$ROUGE - N(s) = \frac{\sum_{r \in R} < \phi_n(r), \phi_n(s) >}{\sum_{r \in R} < \phi_n(r), \phi_n(r) >} \tag{2.6}$$

where <.,.> denotes the inner product of two vectors, $R = \{r_1, ..., r_m\}$ denote a set of reference summaries, s denotes a summary generated automatically by a system, and $\phi_n(d)$ denotes a binary vector contained in a document d. Thus, the metric simply attempts to measure the extent to which the generated summary contains the same information as the reference summaries.

Other ROUGE metrics apply different techniques with the same primary intent. For example, ROUGE-L attempts to find the *longest common subsequence* (LCS) between two summaries, with the rationale that summaries with longer LCSs are more similar. Another ROUGE metric, known as ROUGE-S, is a *gappy* version of the ROUGE-N metric (for n = 2) and matches ordered bi-grams of the generated summary with reference summaries.

The various ROUGE metrics have been shown to have good correlation with human-produced rankings of summarisers. Since the introduction of ROUGE, its popularity has seen its widespread use in evaluating automatic summarisation systems across various domains, including systems for evidence-based medicine, such as the BIOSQUASH and EPoCare systems (as already mentioned).

BLEU

BLEU [Papineni et al., 2002] is an automatic evaluation approach that was originally used for machine translation and has been shown to be promising for automatic summary evaluation as well [Lin and Hovy, 2002]. In this technique, automatically computed accumulative n-gram matching scores (NAMS) between a model unit (MU) and a system summary (S) are used as performance indicators of the system (only content words are used in the n-grams). A number of combinations of n-grams are used to compute NAMS, and the technique is shown to have satisfactory correlation with human scores.

Information-theoretic Evaluation

Lin et al. [2006] introduce an information-theoretic approach to the automatic evaluation of summaries based on the divergence of distribution of terms between an automatic summary and a set of reference summaries. The approach only attempts to evaluate content coverage of the summary through the use of a statistical framework. For a set of documents D, the authors assume that there exists a probabilistic distribution with parameters specified by θ_R that generates reference summaries from D, and the task of summarisation is to estimate θ_R. Similarly, the authors assume that every system summary is generated from a probabilistic distribution with parameters specified by θ_A. The process of summary evaluation then becomes the task of estimating the distance between θ_R and θ_A. The authors present a number of variants of divergence measures (*e.g.*, Jensen-Shannon divergence (JS), Kullback-Leibler divergence (KL)) for this and show that this technique is comparable to ROUGE for the evaluation of single-document summarisation and better than ROUGE in evaluating multi-document summarisation systems.

Louis and Nenkova [2008] also present an evaluation approach based on the same principles. In their approach, they compare the term distributions in the summaries and the original texts using KL and JS divergence, cosine similarity, as well as uni-gram and multinomial models of text. JS divergence is shown to perform best with a high correlation with human measures.

Kabadjov et al. [2010] present yet another information-theoretic evaluation approach that combines an n-gram co-occurrence based measure with a measure of content similarity. The amount of content shared between a pair of texts is measured on the basis of the average semantic similarity between a set of concepts within the first model text and the set of concepts within the second text. Similarity is measured based on Resnik's semantic similarity measure [Resnik, 1995] using the MeSH taxonomy. The approach is capable of capturing complex phenomena such as hypernymy and synonymy and complements the evaluation based on n-gram co-occurrences, resulting in a better and more reliable evaluation metric.

ParaEval

The motivation behind this evaluation technique is the lack of semantic matching of content in automatic evaluation. The authors of this evaluation technique [Zhou et al., 2006] explain that an essential part of semantic matching involves paraphrase matching and this evaluation system attempts to perform that. ParaEval applies a three-level comparison strategy. At the top level, an optimal search via dynamic programming to find multi-word to multi-word paraphrase matches between generated and reference summaries is used. In the second level, a greedy algorithm is used to find single-word paraphrase matches among non-matching fragments in the first level.

Finally, literal lexical uni-gram matching is performed on the remaining text at the third level. The authors show that the quality of ParaEval's evaluations closely resembles that of ROUGE.

2.7.5 Discussion of Evaluation Techniques for Evidence-based Medicine

If summarising for evidence-based medicine is a hard task, evaluating summaries is even harder. A summarisation system for evidence-based medicine should be capable of extracting evidence from medical articles and additionally assess the grade of the evidence. The evaluation should be able to determine if the evidence is correctly extracted and also if the extracted information correctly answers the practitioner's query. Therefore, the single most important aspect of the summary is its content and a strong focus on extrinsic evaluation is required. Intrinsic evaluation to assess aspects such as coherence and readability are perhaps not very important. This is exactly what Sparck Jones [1999] suggested — a stronger focus on the purpose of summaries.

For the automatic evaluation of summaries, approaches based on n-gram co-occurrence such as ROUGE are the most frequently used. Although these approaches are very robust and efficient, their main drawback is that if two summaries were produced using non- or almost non-overlapping vocabulary, yet conveying the same information, the similarity score such summaries would be assigned by purely n-gram based metrics would be too low and, hence, unrepresentative of the actual information they share. This is definitely not desirable for evaluation, particularly in the medical domain where relations such as synonymy and hyponymy play important roles. Furthermore, making relative comparisons between different summarisers is difficult using automatic evaluation approaches such as ROUGE. Thus, approaches that incorporate domain knowledge, semantic similarities, and comparisons between summarisation systems are required. Research on the evaluation of summarisers for evidence-based medicine is still very much in its infancy. However, recent works in this area such as that of Fiszman et al. [2009] have made useful contributions by ensuring that evaluation techniques assess the performance of the summariser in the light of its goals.

2.8 Chapter Conclusion

In this chapter, we provided an overview of evidence-based medicine and described how summarisation of text can aid practitioners at point-of-care. We discussed in detail the obstacles that evidence-based medicine practitioners face, as indicated by various research papers on the topic. Based on the obstacles faced by practitioners, we argued that there are various NLP tasks that can support evidence-based medicine and improve patient care in the long run. Following our review of the domain, we provided an overview of automatic summarisation, its intent and some important contributions to automatic summarisation research. We discussed that unlike automatic

summarisation research on some domains such as news, the medical domain has not received much research attention. We also explained that domain-independent summarisation techniques lack sufficient domain knowledge, incorporation of which can be vital for summarisation research. We reviewed several recent summarisation systems that are customised for the medical domain. Our detailed review of these systems revealed numerous promising approaches, including the clever use of domain-specific information and distributional semantics. Our survey indicates that combining some of the useful approaches from existing literature, and building on from these already explored techniques, may produce encouraging results for text summarisation in this domain. Our survey reveals that an important factor limiting summarisation research in this domain is the lack of suitable corpora/gold standard summaries. Therefore, it is likely that automatic summarisation research in this domain will benefit from the creation of specialised corpora. Furthermore, our survey also revealed that comparing the relative performance of each system is quite impossible because of the unavailability of standard corpora and evaluation criteria.

The survey provided in this chapter sets the context for the research described in this thesis. The survey shows that the first requirement for summarisation research in this specific domain is to analyse the content requirements for evidence-based summaries. Such analysis work requires data that is generated from real life evidence-based medicine practice. In Chapter 3, we describe the corpus we use for our research — a corpus that is specialised for evidence-based medicine summarisation. We describe the design and annotation of the corpus and its possible uses. Through an analysis of the data, we identify the important aspects of evidence-based summarisation and discuss that, in this domain, summarisation involves text compression plus an appraisal of the quality of the evidence.

3 Data

3.1 Introduction

In Chapter 1, we discussed the task of summarisation for evidence-based medicine, and we provided a detailed review of the literature in the areas related to this task in Chapter 2. In the review, we argued that summarisation research in this domain is likely to benefit from the creation of specialised corpora. Our review of specific systems in this domain depicts the weaknesses of existing summarisation systems in the medical domain. It also shows how the creation and evaluation of systems is hindered by the lack of specialised data.

In our research, we utilise a specialised corpus to develop our data-driven summarisation approaches. In this chapter, we introduce this corpus. The corpus was collected semi-automatically from a renowned source of evidence-based medicine information, and edited and annotated manually. In the remainder of this chapter we provide a detailed description of the corpus, including the choice of our data source, data collection techniques, annotations, and the various possible uses of the data. The chapter explains how we use the corpus to analyse the requirements of an evidence-based summariser.

Our target domain and research tasks require a corpus that is very specialised and yet flexible for use in various tasks. Our domain of interest is evidence-based medicine, an important and specialised sub-domain of the broader medical domain. To investigate the tasks associated with evidence-based answer generation, and to explore possible approaches for performing this task, we require data from real life evidence-based medicine practice where question-oriented medical summaries are generated by domain experts. In particular, we require these queries, human-authored answers to clinical queries, and source documents from which the answers have been derived. While there is an abundance of data for automatic summarisation tasks in other domains, to the best of our knowledge there is no corpus suitable for such a specific task, as

explained by Mollá [2010]. There are sources of clinical questions with their answers that could be used as development and evaluation corpora, such as the Parkhurst Exchange collection[1], but to our knowledge, none of the answers in these collections contain explicit pointers to primary literature. There are also other corpora in the medical and biomedical domains that are suited for specific natural language processing tasks, such as the BioScope[2] corpus, which is annotated for negations, speculations, and their linguistic scopes. However, such corpora are not capable of supporting research on automatic medical Question Answering (QA). Also, analysing such corpora does not reveal the specific tasks and sub-tasks associated with evidence-based answer generation. Due to the importance, and the lack of, corpora for biomedical question answering research, the BioASQ challenge[3] recently introduced question-answer data suitable for such research[4].

Because of the unavailability of suitable data for our research, our research group built a corpus using data from the *Clinical Inquiries* section of the Journal of Family Practice[5] (JFP) and the Medline database. The corpus design and creation are described by Mollá [2010], and Mollá and Santiago-Martinez [2011]. The author of this thesis was part of the research group involved in the creation of the corpus, and one of the three annotators of the corpus, as explained later in this chapter. The corpus is suitable for the following tasks:

- Query-focused, single-document summarisation;
- Query-focused, multi-document summarisation;
- Automatic grading quality of evidence;
- Automatic clustering of medical text; and
- Medical Information Retrieval (Possibly).

In this thesis, we utilise the corpus for the first three tasks above. Chapters 4, 5, and 6 describe the details of these tasks. In this chapter, we provide examples from the corpus and verify the suitability of the corpus for these tasks. Based on the structure of the corpus, and our analysis of it, we present a model for our task of evidence-based summarisation. Our model is influenced by the content and structure of the corpus, which contains sample data from real life evidence-based medicine practice. In particular, the specialised data in the corpus reveals that summarisation for evidence-based answer generation involves, in addition to text summarisation, a process for the

[1] http://www.parkhurstexchange.com/searchQA. Accessed on 26th May, 2014.
[2] http://www.inf.u-szeged.hu/rgai/bioscope. Accessed on 26th May, 2014.
[3] http://bioasq.org/. Accessed on 26th May, 2014.
[4] Note that this data set was not available when we commenced our research on evidence-based text summarisation.
[5] http://www.jfponline.com. Accessed on 26th May, 2014.

appraisal of evidence. We show how the task of grading the quality of evidence can be modelled as a text classification problem with three classes. For the single-document and multi-document summarisation tasks, we provide examples from our corpus showing how the summarisation tasks can be modelled and evaluated.

The rest of this chapter is organised as follows: in Section 3.2, we describe the *Clinical Inquiries* section of the JFP, the source from which we obtain the data for our corpus; in Section 3.3, we detail the data collection strategies from both the JFP and the Medline database, and the annotation of the data; in Section 3.4, we provide some important statistics from our corpus and some examples depicting the uses of the corpus; and in Section 3.5, we briefly present the models we propose for the tasks addressed in this thesis.

3.2 The Journal of Family Practice

The Journal of Family Practice is an American medical journal focusing on the family practice domain. Each monthly issue of the journal contains a section called *Clinical Inquiries*, and all the data for our corpus is collected from this section. The data collection process and the annotation is explained by Mollá and Santiago-Martinez [2011]. We now explain the structure and content of this section of the journal, and its applicability to research on automatic evidence-based answer generation.

The *Clinical Inquiries* section of the JFP is a reliable source of information regarding real life evidence-based medicine practice. The articles in this section, being evidence-based answers to clinical queries, are similar in terms of structure. Each article has the same goal: to provide a systematic analysis of the best available medical evidence in response to a clinical query. Thus, preparation of each article in this section requires domain experts to analyse the best available medical evidence, summarise the relevant evidence, and derive bottom-line recommendations.

As a collection, these articles are attractive to work with as they have a number of features amenable to the study of evidence-based summarisation. Importantly, each article contains both single- and multi-document summaries, allowing for the creation of a corpus that can flexibly support research in both forms of text summarisation. Furthermore, the *Clinical Inquiries* articles contain additional information such as grades indicating the qualities of the evidences. This makes it possible for the corpus built from this data source to support research in related but relatively unexplored areas of evidence-based medicine, such as the automatic appraisal of evidence. Data from this section of the JFP also enables the analysis and replication of the various tasks associated with evidence-based medicine practice.

Figures 3.1 and 3.2 in pages 70 and 71 show the important parts of an article from the *Clinical*

Inquiries section of JFP[6]. As the figures show, an article in this section contains:

1. A clinical question. This appears as the title of each *Clinical Inquiry* article. In figure 3.1, the clinical query is: *Which treatments work best for hemorrhoids?*.

2. The bottom-line, evidence-based answer(s) or recommendation(s). A single question may have more than one bottom-line answer, since the question may be answered according to distinct pieces of evidence. The intent of each bottom-line answer is to summarise the evidence from medical literature, associated with the question. Thus, these summaries generally synthesise information from multiple medical publications and state the best possible action to address the medical problem(s) mentioned in the question.

   In the example shown in figure 3.1, there are three bottom-line answers associated with the question. Each of the three bottom-line answers focus on a separate topic. In this case, the first bottom-line answer proposes an intervention for *thrombosed external hemorrhoids*, the second for *prolapsed internal hemorrhoids*, and the third answer provides a summary of *nonoperative techniques*.

3. For each part of the evidence-based answer, a grade indicating the quality of the evidence associated with that answer. This grade is also known as the Strength Of Recommendation (SOR). The purpose of the grade is to indicate the reliability/quality of the synthesised bottom-line answer on the chosen scale. The reliability/quality of a synthesised answer depends on the body of evidence associated with it. A brief justification behind the choice of the grade is generally provided. The presence of evidence grades with each bottom-line answer suggests that appraising the quality of evidence is a vital subtask in evidence-based summary generation.

   In figure 3.1, the first bottom-line answer is given a grade *B*, while the other two answers are assigned *A* grades. The first evidence is obtained from *Retrospective Studies*, which is of lower quality compared to the *Systematic Reviews* from which the second and third answers have been obtained.

4. Detailed justifications for the evidence-based answers. The detailed justifications provide more information about the clinical studies that provide the evidence for the bottom-line answers. Generally, they contain some background information about the studies in addition to the outcomes. The detailed justifications typically contain nuggets of summarised information from individual documents. These summaries are authored by human experts and provide the query-relevant information in a condensed format.

[6]The original article can be found at http://www.jfponline.com/index.php?id=22143&tx_ttnews[tt_news]=174638. Accessed on 26th May, 2014.

In figure 3.1, the detailed justifications are shown in the *Evidence Summary* section. It can be seen from the figure that for each study that is discussed, some background information along with the outcome information is provided. For example, the following extract is taken from a single Retrospective Study, and suggests that *surgical interventions* are recommended over *conservative interventions*.

> *A retrospective study of 231 patients treated conservatively or surgically found that the 48.5% of patients treated surgically had a lower recurrence rate than the conservative group (number needed to treat [NNT]=2 for recurrence at mean follow-up of 7.6 months) and earlier resolution of symptoms (average 3.9 days compared with 24 days for conservative treatment).*

5. References to published medical research papers from which the information in the detailed justifications have been obtained. Thus, these referenced articles are the source texts from which information have been condensed and extracted to produce the detailed justifications.

   Figure 3.2 shows the references of the *Clinical Inquiry* article. From the figure, it can be seen that the detailed justification mentioned above comes from a paper titled *Thrombosed external hemorrhoids: outcome after conservative or surgical management*.

The articles in *Clinical Inquiries* section of the JFP are very carefully constructed reviews. The relevant documents are identified via exhaustive literature searches by medical experts. Mollá and Santiago-Martinez [2011] point out that there are several significant advantages of using data from this resource, rather than direct systematic reviews such as the Cochrane Reviews[7], as a source for this corpus:

1. The format is relatively uniform across all inquiries. Therefore, it enables a semi-automatic method to convert the data to a corpus that can be used by a machine.
2. The text in each inquiry is much more compact than in a Cochrane review, making the target text closer to what a busy healthcare practitioner would want to read.
3. The procedure to find answers in the *Clinical Inquiries* section of the JFP is more methodical than some other sources of evidence-based text.
4. The presence of short answers followed by longer explanations makes this resource suitable for summarisation at various levels.

In the next subsection, we explain the data collection process, the annotation of the data, and the creation of the corpus in detail.

[7] http://www.cochrane.org/cochrane-reviews. Accessed on 26th May, 2014.

Which treatments work best for hemorrhoids?

Evidence-based answer

Excision is the most effective treatment for thrombosed external hemorrhoids (strength of recommendation [SOR]: **B**, retrospective studies). For prolapsed internal hemorrhoids, the best definitive treatment is traditional hemorrhoidectomy (SOR: **A**, systematic reviews). Of nonoperative techniques, rubber band ligation produces the lowest rate of recurrence (SOR: **A**, systematic reviews).

Evidence summary

External hemorrhoids originate below the dentate line and become acutely painful with thrombosis. They can cause perianal pruritus and excoriation because of interference with perianal hygiene. Internal hemorrhoids become symptomatic when they bleed or prolapse (TABLE).

For thrombosed external hemorrhoids, surgery works best

Few studies have evaluated the best treatment for thrombosed external hemorrhoids. A retrospective study of 231 patients treated conservatively or surgically found that the 48.5% of patients treated surgically had a lower recurrence rate than the conservative group (number needed to treat [NNT]=2 for recurrence at mean follow-up of 7.6 months) and earlier resolution of symptoms (average 3.9 days compared with 24 days for conservative treatment).[1]

Another retrospective analysis of 340 patients who underwent outpatient excision of thrombosed external hemorrhoids under local anesthesia reported a low recurrence rate of 6.5% at a mean follow-up of 17.3 months.[2]

A prospective, randomized controlled trial (RCT) of 98 patients treated nonsurgically found improved pain relief with a combination of topical nifedipine 0.3% and lidocaine 1.5% compared with lidocaine alone. The NNT for complete pain relief at 7 days was 3.[3]

Conventional hemorrhoidectomy beats stapling

Many studies have evaluated the best treatment for prolapsed hemorrhoids. A Cochrane systematic review of 12 RCTs that compared conventional hemorrhoidectomy with stapled hemorrhoidectomy in patients with grades I to III hemorrhoids found a lower rate of recurrence (follow-up ranged from 6 to 39 months) in patients who had conventional hemorrhoidectomy (NNT=14).[4] Conventional hemorrhoidectomy showed a nonsignificant trend in decreased bleeding and decreased incontinence.

A second systematic review of 25 studies, including some that were of

Figure 3.1: Extract from a sample article in the *Clinical Inquiries* section of the Journal of Family Practice showing the title, bottom-line summary and detailed justifications.

Nonoperative techniques? Consider rubber band ligation

A systematic review of 3 poor-quality trials comparing rubber band ligation with excisional hemorrhoidectomy in patients with grade III hemorrhoids found that excisional hemorrhoidectomy produced better long-term symptom control but more immediate postoperative complications of anal stenosis and hemorrhage.[6] Rubber band ligation had the lowest recurrence rate at 12 months compared with the other nonoperative techniques of sclerotherapy and infrared coagulation.[7]

Fiber supplements help relieve symptoms

A Cochrane systematic review of 7 RCTs enrolling a total of 378 patients with grade I to III hemorrhoids evaluated the effect of fiber supplements on pain, itching, and bleeding. Persistent hemorrhoid symptoms decreased by 53% in the group receiving fiber.[8]

When surgical hemorrhoidectomy is recommended

The American Society of Colon and Rectal Surgeons recommends adequate fluid and fiber intake for all patients with symptomatic hemorrhoids. For grade I to III hemorrhoids, the society states that banding is usually most effective. When office treatments fail, the society recommends surgical hemorrhoidectomy (SOR: **B**).

The society recommends excision of thrombosed hemorrhoids less than 72 hours old and expectant treatment with analgesia and sitz baths for thrombosed hemorrhoids present for longer than 72 hours (SOR: **B**).[9]

The American Gastroenterological Association recommends excision

I	Hemorrhoids do not prot
II	Hemorrhoids protrude w reduce spontaneously
III	Hemorrhoids protrude ar by hand
IV	Hemorrhoids are permar

Source: Madoff RD, et al. *Gastroenterology.* 2004.[10]

hemorrhoids that present early. Surgical hemorrhoidectomy should be reserved for when conservative treatment fails and for patients with symptomatic grade III and IV hemorrhoids.[10] ■

References

1. Greenspon J, Williams SB, Young HA, et al. Thrombosed external hemorrhoids: outcome after conservative or surgical management. *Dis Colon Rectum.* 2004;47:1493-1498.
2. Jongen J, Bach S, Stubinger SH, et al. Excision of thrombosed external hemorrhoids under local anesthesia: a retrospective evaluation of 340 patients. *Dis Colon Rectum.* 2003;46:1226-1231.
3. Perrotti P, Antropoli C, Molino D, et al. Conservative treatment of acute thrombosed external hemorrhoids with topical nifedipine. *Dis Colon Rectum.* 2001;44:405-409.
4. Jayaraman S, Colquhoun PH, Malthaner RA. Stapled versus conventional surgery for hemorrhoids. *Cochrane Database Syst Rev.* 2006;(4):CD005393.
5. Tjandra JJ, Chan MK. Systematic review on the procedure for prolapse and hemorrhoids (stapled hemorrhoidopexy). *Dis Colon Rectum.* 2007;50:878-892.
6. Shanmugam V, Thaha MA, Rabindranath KS, et al. Systematic review of randomized trials comparing rubber band ligation with excisional haemorrhoidectomy. *Br J Surg.* 2005;92:1481-1487.
7. Johanson JF, Rimm A. Optimal nonsurgical treatment of hemorrhoids: a comparative analysis of infrared coagulation, rubber band ligation, and injection sclerotherapy. *Am J Gastroenterol.* 1992;87:1600-1606.
8. Alonso-Coello P, Guyatt G, Heels-Ansdell D, et al. Laxatives for the treatment of hemorrhoids. *Cochrane Database Syst Rev.* 2005(4):CD004649.
9. Cataldo P, Ellis CN, Gregorcyk S, et al. Practice parameters for the management of hemorrhoids (revised). *Dis Colon Rectum.* 2005;48:189-194.
10. Madoff RD, Fleshman JW, American Gastroenterological Association Clinical Practice Committee. American Gastroenterological Association technical review on the diagnosis and treatment of hem-

Figure 3.2: Extract from a sample article in the *Clinical Inquiries* section of the Journal of Family Practice showing the detailed justifications and references at the bottom of the article.

3.3 Data Collection and the Corpus

The collection of data from JFP to produce the corpus involved several steps namely: automatic extraction and conversion of text, and manual annotations. In this section, we first provide an overview of the corpus itself and then discuss some of the data collection details. We commence by providing an overview of the corpus using formal notations that we reuse later on in the chapter.

3.3.1 Corpus Overview

The corpus is encoded in XML format and consists of a set of records, $R = \{r_1 \ldots r_m\}$. Each record, r_i, represents a *Clinical Inquiry* article from the JFP, and contains one clinical query, q_i, so that we have a set of questions $Q = \{q_1 \ldots q_m\}$. Each r_i has associated with it a set of one or more bottom-line answers to the query, $A_i = \{a_{i1} \ldots a_{in}\}$. Each bottom-line answer has a grade, g_{ij}, associated with it, so that each record contains a set of grades, $G_i = \{g_{i1} \ldots g_{in}\}$. For each bottom-line answer of r_i, a_{ij}, there exists a set of human-authored detailed justifications (single-document summaries) $L_{ij} = \{l_{ij1} \ldots l_{ijo}\}$. Each detailed justification, l_{ijk}, is associated with at least one source document abstract d_{ijk}. Thus, the corpus has a set of source documents, which we denote as $D_{ij} = \{d_{ij1} \ldots d_{ijo}\}$.[8] Figure 3.3 diagrammatically illustrates the structure of a sample record from the corpus. In the figure, there are two bottom-line summaries associated with the question (each with a grade indicating the quality of evidence). Each bottom-line summary is associated with two detailed justifications, which in turn are associated with one source text each. This structure of the corpus was chosen because it is ideal for both single- and multi-document, query-focused, text-to-text summarisation.

We now briefly look back at how the data for the corpus was collected.

3.3.2 Data Extraction

Data were collected from all the publicly available *Clinical Inquiries* articles of the JFP (dating from the year 2001 to 2010), after obtaining permission from the publishers. The questions, bottom-line answers, evidence grades, detailed justification texts and references were all automatically downloaded and stored in a database.

[8] Note that if there are more than one source documents for a justification, the number o in d_{ijo} should be greater than or equal to the number empho in l_{ijo}.

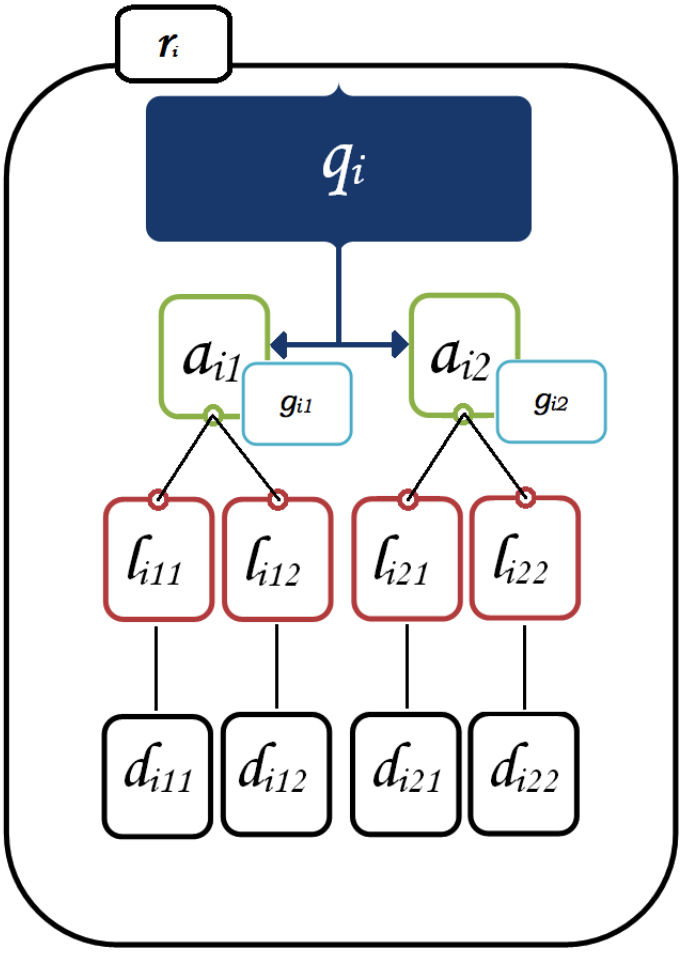


Figure 3.3: Structure of a sample record from the corpus.

3.3.3 Annotation of Detailed Justifications

The source text from the JFP articles did not provide explicit connections from each justification to the specific bottom-line answer, as can be seen from the extract shown in Figure 3.1. Therefore, the corpus had to be prepared by manually identifying the detailed justifications associated with the bottom-line summaries. We utilised a web-based annotation tool[9] that, for each article, displays the question and each of the answer parts. Each bottom-line summary has associated empty text areas where the annotator could copy and paste all the detailed justifications associated with that bottom-line summary.

The total number of pages to annotate was distributed among three annotators, one of whom is the author of this thesis. A small percentage of the pages was annotated by all annotators (the

[9]The annotation tool was designed by a research programmer. Details are provided in Mollá and Santiago-Martinez [2011].

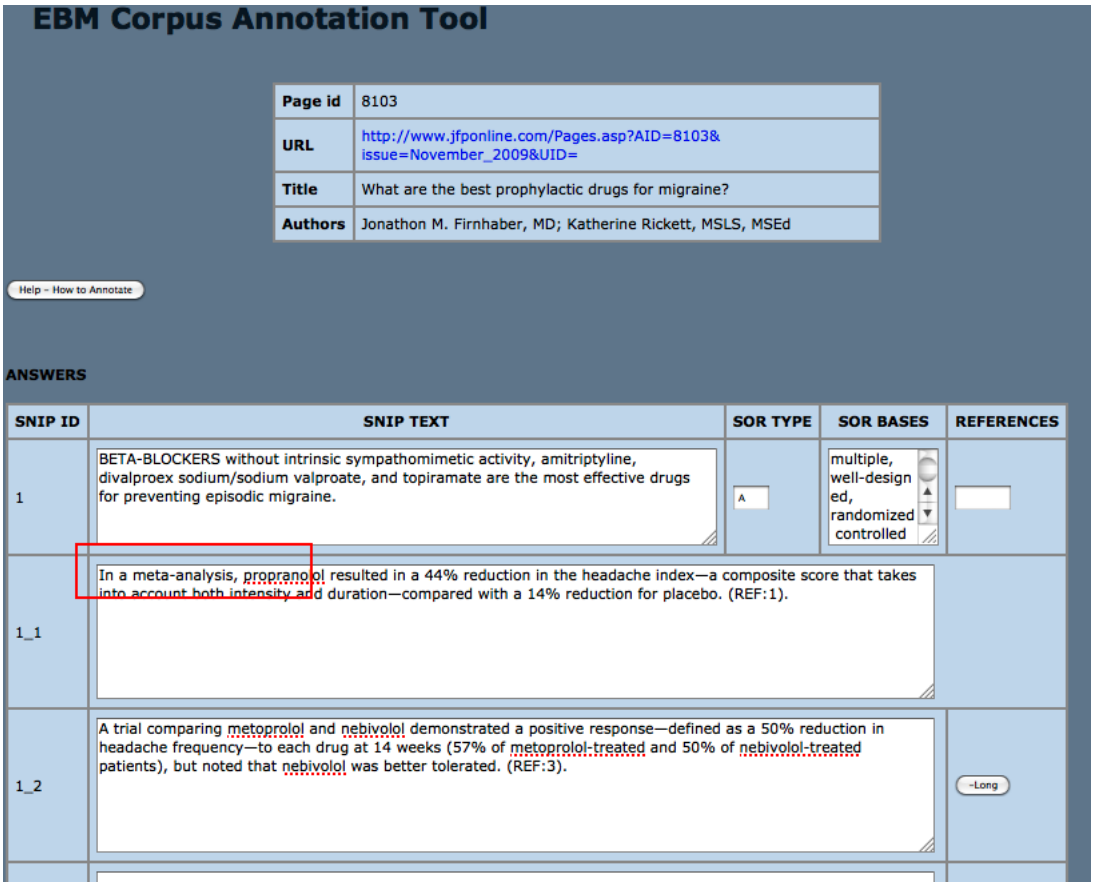


Figure 3.4: Screenshot 1 of the annotation tool.

annotators did not know beforehand which of the pages were annotated by all), to check for inconsistencies. The annotation process was done in several stages, with periodic checks on the common pages to detect and solve systematic inconsistencies in the annotation criteria. During those checks the annotators agreed on a set of criteria, an extract of which is:

1. Remove phrases connecting to text outside the answer justification and modify anaphora to make the text self-contained. For example, change 'In another study' to 'In a study' or 'The second study' to 'A study'. An example of this is provided in the figures 3.4 and 3.5 (the relevant text is indicated by the red squares in the figures). The phrase 'In a 1991 meta-analysis' (shown in figure 3.5) is added as detailed justification 1_1 (shown in figure 3.4) after removing the term '1991'.

2. Remove general text not directly associated with a referenced document or a bottom-line answer.

3. If there are multiple references associated with a justification, split it into separate justifications whenever possible. In the process, some of the text may need to be copied so that

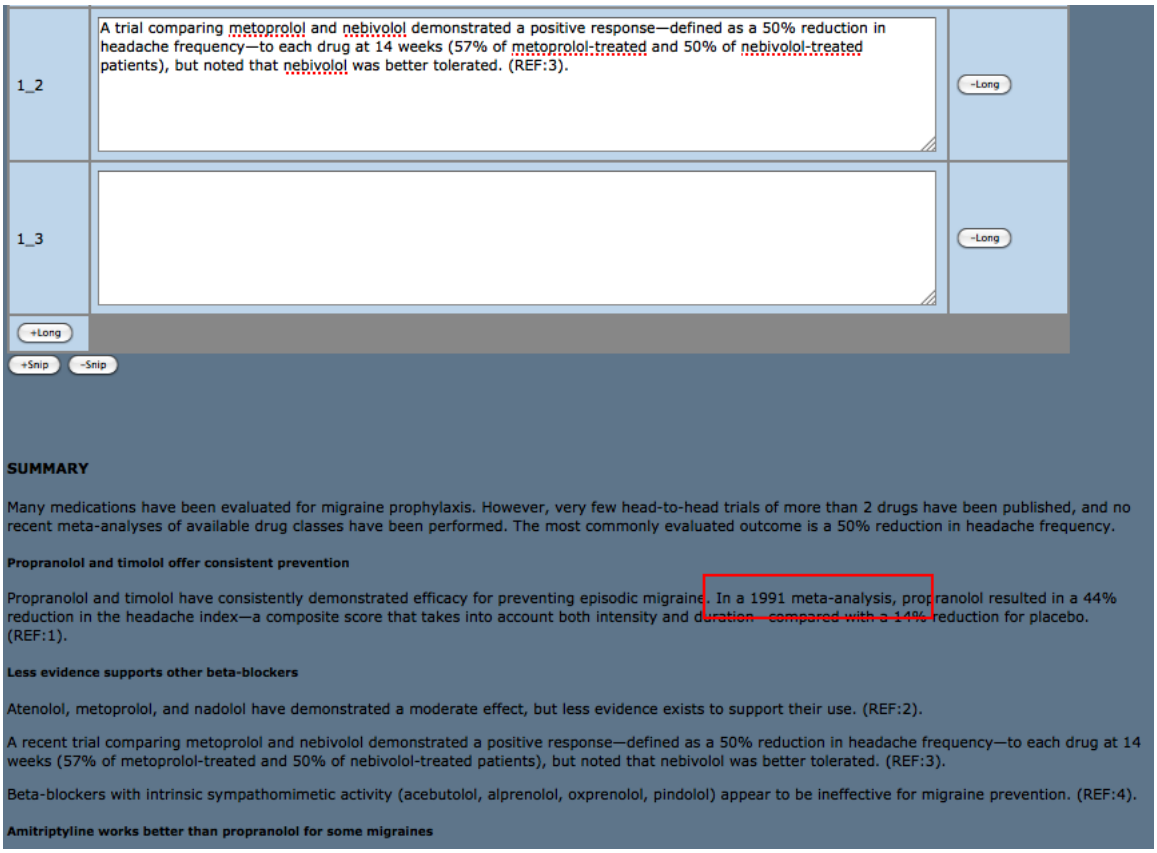
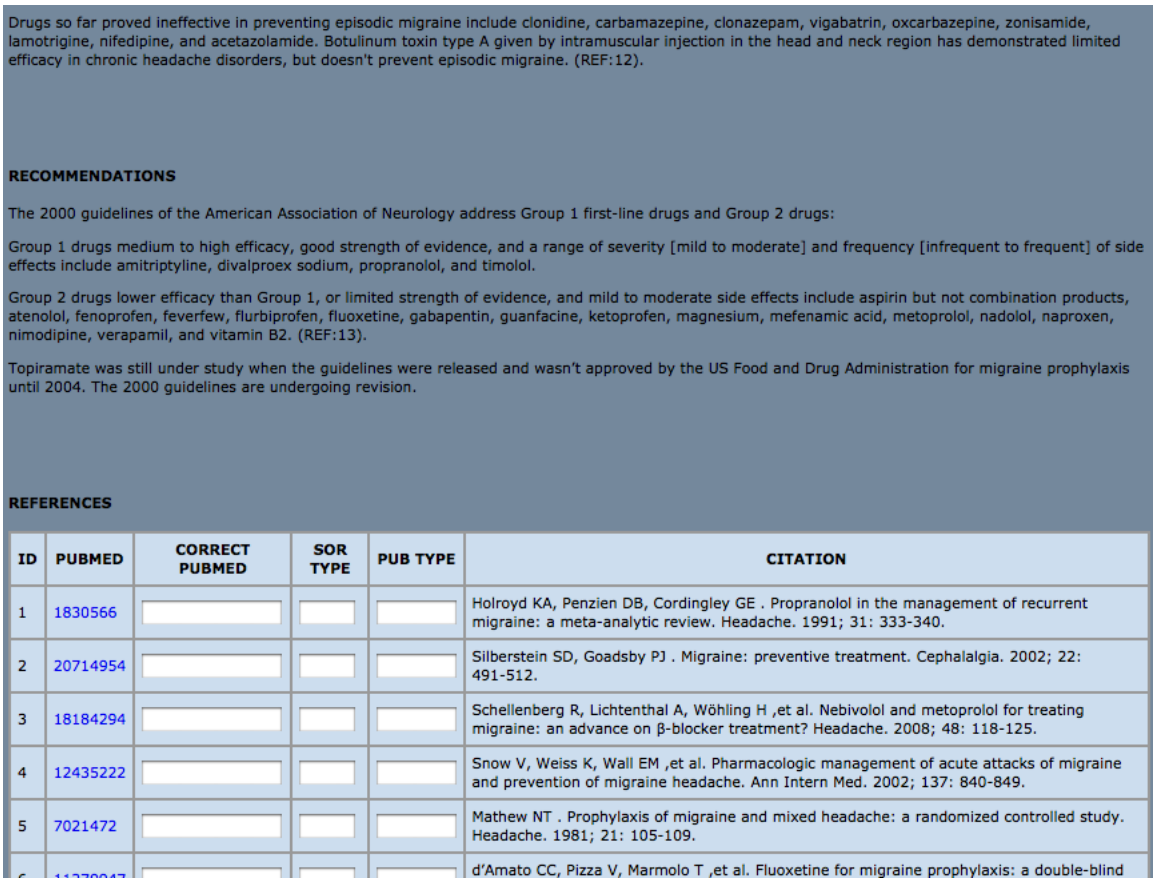


Figure 3.5: Screenshot 2 of the annotation tool.

each justification is self-contained.

4. If a paragraph directly associated to a bottom-line answer does not have any references, manually check if it can be added to the previous or the next paragraph which contains a reference.

These criteria mostly addressed the need for each answer justification to be self-contained, and to match an answer justification to one reference only whenever possible. After inspection of a random sample of the common pages, the annotators agreed that the variations in the annotations were acceptable. Figures 3.4, 3.5, and 3.6 show three screenshots of the annotation tool. The sample record is different to the one shown in Figure 3.1. Figure 3.4 shows the question, the bottom-line answer along with two manually edited detailed justifications, each of which refer to a specific source document from which the information was obtained. Figure 3.5 is the next screenshot showing available text area for insertion of more detailed justifications and some text from the JFP article. Figure 3.6 shows the bottom of the annotation page with the references information, where annotators have the option of correcting errors with PubMed IDs or SORs before saving their work.



Drugs so far proved ineffective in preventing episodic migraine include clonidine, carbamazepine, clonazepam, vigabatrin, oxcarbazepine, zonisamide, lamotrigine, nifedipine, and acetazolamide. Botulinum toxin type A given by intramuscular injection in the head and neck region has demonstrated limited efficacy in chronic headache disorders, but doesn't prevent episodic migraine. (REF:12).

RECOMMENDATIONS

The 2000 guidelines of the American Association of Neurology address Group 1 first-line drugs and Group 2 drugs:

Group 1 drugs medium to high efficacy, good strength of evidence, and a range of severity [mild to moderate] and frequency [infrequent to frequent] of side effects include amitriptyline, divalproex sodium, propranolol, and timolol.

Group 2 drugs lower efficacy than Group 1, or limited strength of evidence, and mild to moderate side effects include aspirin but not combination products, atenolol, fenoprofen, feverfew, flurbiprofen, fluoxetine, gabapentin, guanfacine, ketoprofen, magnesium, mefenamic acid, metoprolol, nadolol, naproxen, nimodipine, verapamil, and vitamin B2. (REF:13).

Topiramate was still under study when the guidelines were released and wasn't approved by the US Food and Drug Administration for migraine prophylaxis until 2004. The 2000 guidelines are undergoing revision.

REFERENCES

ID	PUBMED	CORRECT PUBMED	SOR TYPE	PUB TYPE	CITATION
1	1830566				Holroyd KA, Penzien DB, Cordingley GE . Propranolol in the management of recurrent migraine: a meta-analytic review. Headache. 1991; 31: 333-340.
2	20714954				Silberstein SD, Goadsby PJ . Migraine: preventive treatment. Cephalalgia. 2002; 22: 491-512.
3	18184294				Schellenberg R, Lichtenthal A, Wöhling H ,et al. Nebivolol and metoprolol for treating migraine: an advance on β-blocker treatment? Headache. 2008; 48: 118-125.
4	12435222				Snow V, Weiss K, Wall EM ,et al. Pharmacologic management of acute attacks of migraine and prevention of migraine headache. Ann Intern Med. 2002; 137: 840-849.
5	7021472				Mathew NT . Prophylaxis of migraine and mixed headache: a randomized controlled study. Headache. 1981; 21: 105-109.
6	11279047				d'Amato CC, Pizza V, Marmolo T ,et al. Fluoxetine for migraine prophylaxis: a double-blind

Figure 3.6: Screenshot 3 of the annotation tool.

3.3.4 Extracting Reference Information via Crowdsourcing

In addition to all the information available from the JFP, our target task of summarisation requires the source texts associated with the detailed justifications. Each *Clinical Inquiries* article contains multiple bibliographical references to medical publications. Although a small number of those point to online sources, the majority are full bibliographical references pointing to traditionally published resources such as medical journals or magazines. We wanted to obtain as many of the associated source articles as possible. Since most medical articles are regularly logged into the PubMed database, we chose to obtain electronic copies of the article abstracts from there. It was not possible to obtain full articles since we did not have access to all the venues from which they were available. The easiest way to obtain an electronic copy of an article abstract is to download it from the PubMed website using the article's PubMed ID. Therefore, completing the corpus required identifying the PubMed IDs of the abstracts.

We used crowd-sourcing to obtain as many of the associated abstracts as possible from PubMed.

```
12446     <record id="8103">
12447         <url>http://www.jfponline.com/Pages.asp?AID=8103&amp;issue=November_2009&amp;UID=</url>
12448         <question>What are the best prophylactic drugs for migraine?</question>
12449         <answer>
12450             <snip id="1">
12451                 <sniptext>BETA-BLOCKERS without intrinsic sympathomimetic activity, amitriptyline,
    divalproex sodium/sodium valproate, and topiramate are the most effective drugs for preventing episodic
    migraine.</sniptext>
12452                 <sor type="A">multiple, well-designed, randomized controlled trials [RCTs]</sor>
12453                 <long id="1_7">
12454                     <longtext>A study of patients with migraine without aura found a 57% reduction in total
    pain index-a value based on pain intensity and hours of headache per month-with fluoxetine compared with an
    insignificant 31% reduction with placebo.</longtext>
12455                     <ref id="11279947" abstract="Abstracts/11279947.xml">d&apos;Amato CC, Pizza V, Marmolo
    T ,et al. Fluoxetine for migraine prophylaxis: a double-blind trial. Headache. 1999; 39: 716-719.</ref>
12456                 </long>
12457                 <long id="1_6">
12458                     <longtext>No evidence from controlled trials supports the use of fluvoxamine,
    paroxetine, sertraline, phenelzine, venlafaxine, mirtazapine, trazodone, or bupropion. Initial studies of
    delayed-release divalproex at doses ranging from 500 to 1500 mg daily found that 44% of divalproex-treated
    patients reported a 50% reduction in migraine frequency, compared with 21% in the placebo group (number
    needed to treat [NNT]=4).</longtext>
12459                     <ref id="9137847" abstract="Abstracts/9137847.xml">Klapper J . Divalproex sodium in
    migraine prophylaxis: a dose-controlled study [published correction appears in Cephalalgia. 1997;17:798].
    Cephalalgia. 1997; 17: 103-108.</ref>
12460                 </long>
12461                 <long id="1_5">
12462                     <longtext>A trial found amitriptyline to be more effective than propranolol in mixed
    migraine-tension-type headache, whereas propranolol was more effective for migraine alone.</longtext>
12463                     <ref id="7021472" abstract="Abstracts/7021472.xml">Mathew NT . Prophylaxis of migraine
    and mixed headache: a randomized controlled study. Headache. 1981; 21: 105-109.</ref>
12464                 </long>
12465                 <long id="1_4">
12466                     <longtext>Beta-blockers with intrinsic sympathomimetic activity (acebutolol,
    alprenolol, oxprenolol, pindolol) appear to be ineffective for migraine prevention. Divalproex sodium and
    sodium valproate show strong, consistent evidence of efficacy; they may be particularly useful for patients
    with prolonged or atypical migraine aura.</longtext>
12467                     <ref id="12435222" abstract="Abstracts/12435222.xml">Snow V, Weiss K, Wall EM ,et al.
    Pharmacologic management of acute attacks of migraine and prevention of migraine headache. Ann Intern Med.
    2002; 137: 840-849.</ref>
12468                 </long>
12469                 <long id="1_3">
12470                     <longtext>A trial comparing metoprolol and nebivolol demonstrated a positive response-
    defined as a 50% reduction in headache frequency-to each drug at 14 weeks (57% of metoprolol-treated and
    50% of nebivolol-treated patients), but noted that nebivolol was better tolerated.</longtext>
12471                     <ref id="18184294" abstract="Abstracts/18184294.xml">Schellenberg R, Lichtenthal A,
    WÖhling H ,et al. Nebivolol and metoprolol for treating migraine: an advance on β-blocker treatment?
    Headache. 2008; 48: 118-125.</ref>
12472                 </long>
12473                 <long id="1_2">
12474                     <longtext>Atenolol, metoprolol, and nadolol have demonstrated a moderate effect, but
    less evidence exists to support their use.</longtext>
12475                     <ref id="12230591" abstract="Abstracts/12230591.xml">Silberstein SD, Goadsby PJ .
    Migraine: preventive treatment. Cephalalgia. 2002; 22: 491-512.</ref>
12476                 </long>
12477                 <long id="1_1">
```

Figure 3.7: Screenshot of a sample record from the corpus.

The crowdsourcing task was created using Amazon Mechanical Turk[10], and the task of the *turkers* was to manually identify the PubMed IDs for the articles referenced in the JFP. Detailed descriptions about the design of this task is provided by Mollá [2010] and Mollá and Santiago-Martinez [2011]. The final accuracy of the annotation task was manually checked on a random sample of 100 references by double checking them. No errors were detected. Finally, once all IDs were found, the abstracts were automatically downloaded from PubMed and added to the corpus. We chose to download the articles in XML format, which contains useful meta-data, the abstract text and additional annotations such as classification tags and MeSH terms. Note that it was not possible to obtain all the abstracts. The next section discusses various important statistics

[10]https://www.mturk.com/mturk/. Accessed on 26th May, 2014.

```
60          </month>
61         </pubdate>
62        </journalissue>
63        <title>
64         Cephalalgia : an international journal of headache
65        </title>
66        <isoabbreviation>
67         Cephalalgia
68        </isoabbreviation>
69       </journal>
70       <articletitle>
71        Migraine: preventive treatment.
72       </articletitle>
73       <pagination>
74        <medlinepgn>
75         491-512
76        </medlinepgn>
77       </pagination>
78       <abstract>
79        <abstracttext>
80         Migraine is a common episodic headache disorder. A
   comprehensive headache treatment plan includes acute attack treatment
   to relieve pain and impairment and long-term preventive therapy to
   reduce attack frequency, severity, and duration. Circumstances that
   might warrant preventive treatment include: (i) migraine that
   significantly interferes with the patient's daily routine despite
   acute treatment; (ii) failure, contraindication to, or troublesome
   side-effects from acute medications; (iii) overuse of acute
   medications; (iv) special circumstances, such as hemiplegic migraine;
   (v) very frequent headaches (more than two a week); or (vi) patient
   preference. Start the drug at a low dose. Give each treatment an
   adequate trial. Avoid interfering, overused, and contraindicated
   drugs. Re-evaluate therapy. Be sure that a woman of childbearing
   potential is aware of any potential risks. Involve patients in their
   care to maximize compliance. Consider co-morbidity. Choose a drug
   based on its proven efficacy, the patient's preferences and headache
   profile, the drug's side-effects, and the presence or absence of
   coexisting or co-morbid disease. Drugs that have documented high
   efficacy and mild to moderate adverse events (AEs) include beta-
   blockers, amitriptyline, and divalproex. Drugs that have lower
   documented efficacy and mild to moderate AEs include selective
   serotonin reuptake inhibitors (SSRIs), calcium channel antagonists,
   gabapentin, topiramate, riboflavin, and non-steroidal anti-
   inflammatory drugs.
81        </abstracttext>
82       </abstract>
83       <affiliation>
84        Jefferson Headache Center, and Thomas Jefferson University
   Hospital, Philadelphia, PA 19107, USA.
   Stephen.Silberstein@mail.tju.ed
85       </affiliation>
86       <authorlist completeyn="Y">
87        <author validyn="Y">
88         <lastname>
89          Silberstein
90         </lastname>
91         <forename>
92          S D
```

Figure 3.8: Screenshot of a sample PubMed abstract from the corpus.

associated with the corpus in more detail.[11]

Figure 3.7 shows the actual corpus in XML format. Figure 3.8 presents a screenshot showing the structure of a PubMed abstract. The abstract (PubMed ID: 12230591) is also chosen from the same record. A sample of a full abstract is provided in Appendix A.

3.4 Statistics and Use of Corpus

The final statistics of the corpus are as follows:

- Number of *records*: 456;
- Number of *bottom-line answers* associated with the records: 1,396;
- Number of *bottom-line answers* with quality grades specified: 1,225 (171 answers did not have any associated quality grades);
- Number of *detailed justifications* associated with the bottom-line answers: 3,036; and
- Number of unique referenced articles (PubMed abstracts) associated with the detailed justifications: 2,908; 797 referenced articles are associated with more than one detailed justification.

On average, each record has 3.06 bottom-line answers; each bottom-line answer has 2.17 associated detailed justifications; each detailed justification has 1.22 references. There are 6.57 references, on average, for each question. Despite our best efforts, we failed to locate a number of the source documents on Medline. All such cases were noted and recorded in the corpus. In total, 312 (10.7%) referenced documents could not be found on Medline. Furthermore, in a number of cases, although the referenced documents were found in Medline, the XML versions of the documents did not contain any text from the abstract. There were a total of 311 (10.7%) such cases.

The distribution of the quality grades is: 345 for A, 535 for B, 330 for C, 15 for D[12]. A total of 171 bottom-line summaries do not have any quality grade assigned, and therefore, these entries are unusable for research in automatic grading of evidence. However, their applicability to automatic summarisation research does not change.

Figures 3.9 and 3.10 illustrate the distributions of bottom-line summaries, detailed justifications, references, and quality grades in our corpus.

[11] The author of this thesis was not involved in the crowd-sourcing experiments.

[12] SORT only has the grades A, B, and C, but some authors used the grade D to specify very poor evidence.

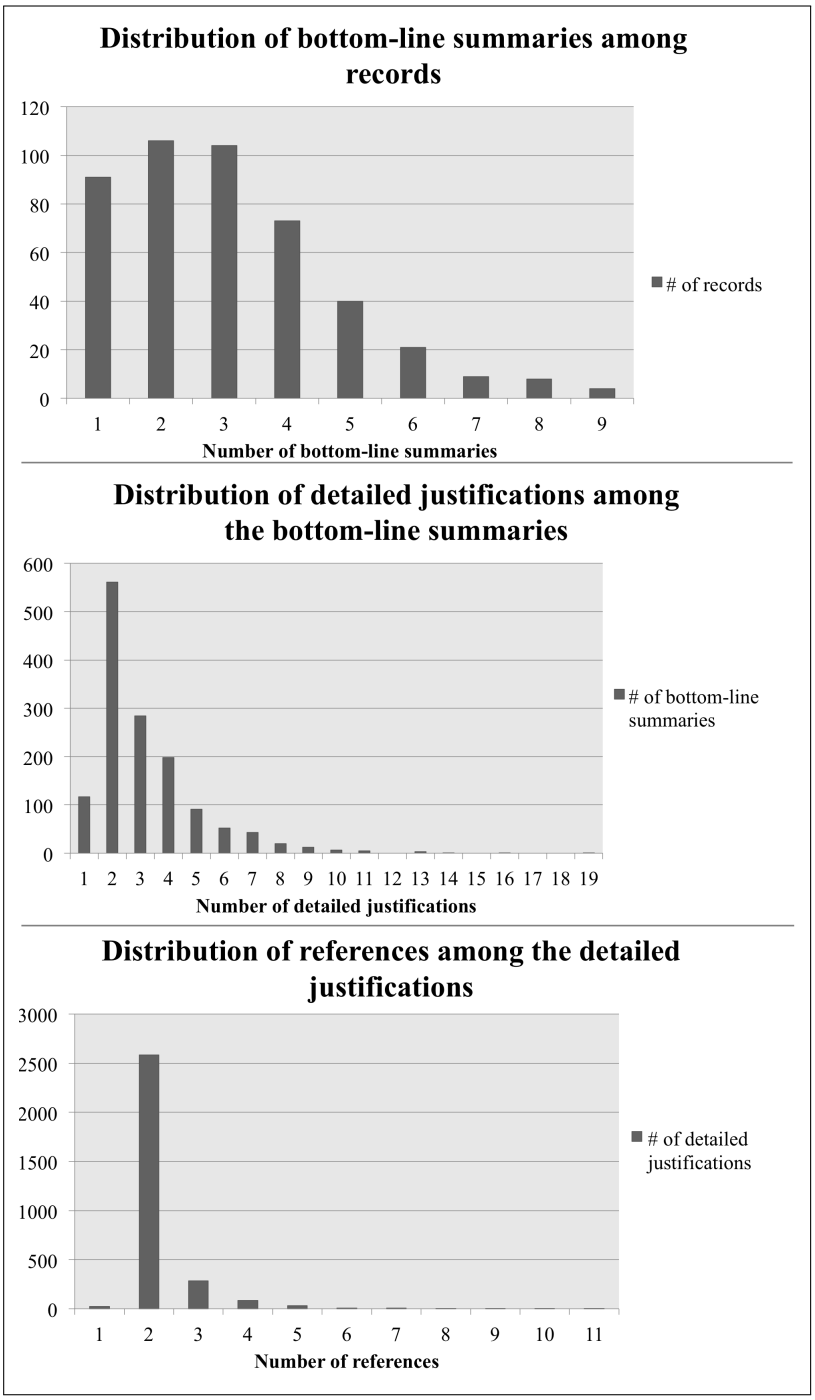


Figure 3.9: Distributions of bottom-line summaries, detailed justifications, and references in our corpus.

From the perspective of this thesis, the corpus has two primary uses: text summarisation and evidence quality appraisal. The tasks and the suitability of the corpus for them are discussed

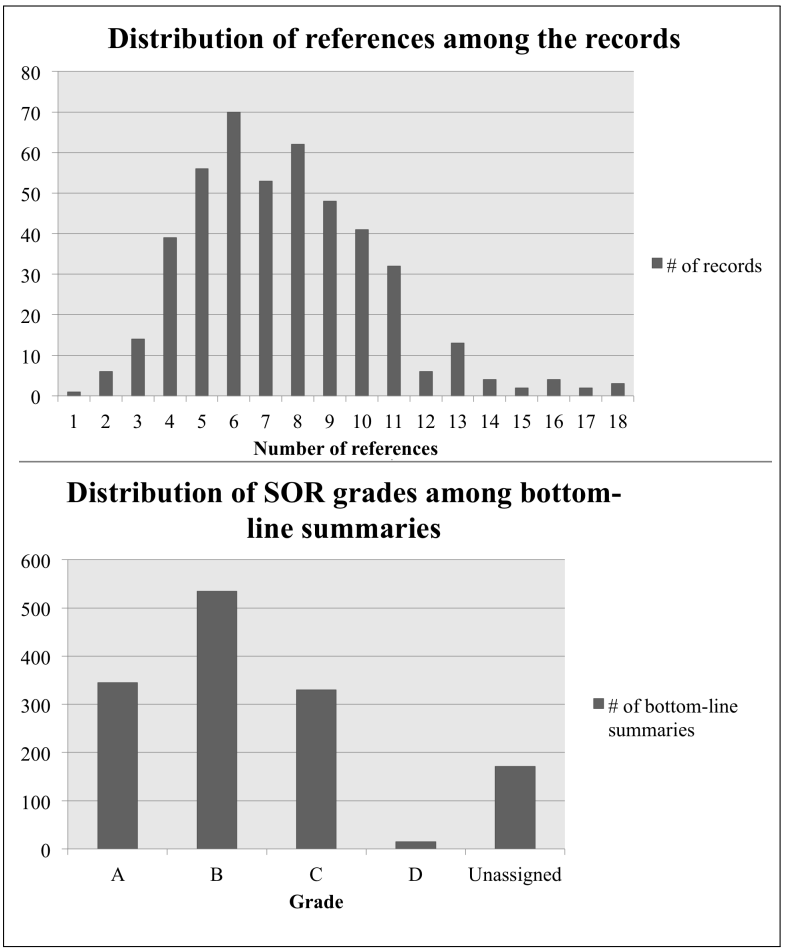


Figure 3.10: Distributions of references and quality grades in our corpus.

below:

Evidence-based summarisation.
The intended primary use of the corpus is to perform automatic, evidence-based text summarisation. The design of the corpus enables the development and testing of single-document text summarisation. For this task, the abstracts can be used as source texts and the human-authored detailed justifications can be used as the gold standard. Since the questions are also available in the corpus, they can be used for performing query-focused, single-document summarisation.

Figure 3.11 shows an example of information in our corpus that can be used for single-document summarisation. The figure shows the question, the human-authored detailed justification (to the bottom-line answer which is not shown in the example), and text from the referenced abstract. The target of automatic summarisation in this case is to identify and extract information from the source texts that most closely resembles the human-authored summaries.

Question.
What are the best prophylactic drugs for migraine?

Detailed justification (single-document summary):
A study of the extended-release form of divalproex sodium demonstrated a 4-week reduction in headache rate to 1.2 from a baseline of 4.4, compared with a decrease of 0.6 for placebo (95% confidence interval [CI] of treatment difference, 0.2-1.2).

[*PubMed ID: 12058094*]

Source text (abstract):

OBJECTIVE: To evaluate the efficacy and safety of extended-release divalproex sodium compared with placebo in prophylactic monotherapy treatment of migraine headache.

METHODS: This was a double-blind, randomized, placebo-controlled, parallel-group study. Subjects with more than two migraine headache attacks during a 4-week baseline were randomly assigned in a 1:1 ratio at each center to receive either extended-release divalproex sodium or matching placebo once daily for 12 weeks. Subjects initiated treatment on 500 mg once daily for 1 week, and the dose was then increased to 1,000 mg once daily with an option, if intolerance occurred, to permanently decrease the dose to 500 mg during the second week. Reduction from baseline in 4-week migraine headache rate was the primary efficacy variable. Migraine headaches separated by a less than 24-hour headache-free interval were counted as single migraines in calculating migraine headache rates. Tolerance and safety were also evaluated.

RESULTS: The mean reductions in 4-week migraine headache rate were 1.2 (from a baseline mean of 4.4) in the extended-release divalproex sodium group and 0.6 (from a baseline mean of 4.2) in the placebo group (p = 0.006); reductions with extended-release divalproex sodium were significantly greater than with placebo in all three 4-week segments of the treatment period. No significant differences were detected between treatment groups in either the overall incidence or in the incidence of any specific treatment-emergent adverse event; 8% of subjects treated with extended-release divalproex sodium and 9% of those treated with placebo discontinued for adverse events.

CONCLUSIONS: Extended-release divalproex sodium is an efficacious, well-tolerated, safe, and easy-to-use once-a-day prophylactic antimigraine medication.

Figure 3.11: An example of single-document summarisation from our corpus, showing the question, the summary and the source abstract. Note that as can be seen from the example, not all the information contained in the detailed justification is contained in the abstract. It is likely that this information originates from the full text of the article.

The corpus is also suitable for performing evidence-based multi-document summarisation. For this task, the bottom-line answers are the target summaries, and all the abstracts associated with the bottom-line summary are the source texts. Figure 3.12 provides an example from our corpus. The bottom-line summary is associated with two detailed justifications (not shown in the example), which in turn cite two research publications. The target of the summarisation task is therefore to synthesise information from the source texts to generate the bottom-line summaries.

Evidence Appraisal.
The literature associated with evidence-based medicine suggests that the appraisal of evidence is an integral component of evidence-based decision making. It is also a time-consuming component of the medical practice. Data in our corpus suggests that the appraisal of evidence can be considered to be an important subtask of the overall summarisation process. However, there has been minimal research on the systematic quality assessment of medical evidence. Our corpus contains annotations for the qualities of the evidences associated with the bottom-line answers. The grades assigned use the Strength of Recommendation (SOR) taxonomy [Ebell et al., 2004] and the grade assignment process involves the systematic assessment of the source documents to identify key information indicating the evidence grade. As explained earlier, the quality grade for a bottom-line recommendation depends on the evidence associated with it. Our corpus contains the evidence associated with each recommendation in the form of the PubMed articles (d_{ijk}). We, therefore, use information from these articles to predict evidence quality grades for the bottom-line recommendations.

Figure 3.13 gives an example of the type of information in our corpus that can be used for the automatic grading task. The figure shows the question, the grade assigned to a bottom-line summary associated with the question (the bottom-line summary is not shown in this case), and the source abstracts responsible for the grade assigned. The target of this task is therefore to analyse the information in the source texts to deduce the quality of the evidence presented in them. Note that the input information for this task may be identical to the summarisation task. Because their inputs are the same, in our evidence-based summarisation model, the two tasks of evidence appraisal and text summarisation are executed in parallel. We model the problem of automatic grading of evidence as a text classification task and attempt to solve it using machine learning algorithms.

3.5 A Model for Automatic, Evidence-based Summarisation

Now that we have explained the data contained in our corpus, we formalise the tasks associated with evidence-based summarisation and provide an overview of our summarisation model. Based on our analysis of the corpus, we divide the task of evidence-based summarisation to generate

Question.
What are the best therapies for acute migraine in pregnancy?

Bottom-line answer (multi-document summary):
Three treatment studies suggest that nonpharmacological therapies (combinations of skin warming, relaxation, biofeedback, and physical therapy) were effective for pain relief. [*PubMed IDs: 8600478, 2401622*]

Source text 1 [PMID: 8600478]:
Concerns about the effects of maternal medications on the growing baby limit the use of medication treatment for benign conditions, such as recurring headaches, during pregnancy and lactation. Nonpharmacological therapies hold particular promise for pregnant women due to the limited medication options. No controlled studies, however, have reported on the efficacy of nonpharmacological treatments for pregnant women. The first study evaluated the effectiveness of a combined nonpharmacological treatment (CT) consisting of relaxation, skin-warming biofeedback, and physical therapy for pregnant women with chronic headaches. In a second study, the CT protocol was compared with an attention control (AC) that received headache education and skin-cooling biofeedback. The first study resulted in significant symptom improvement in 79% of subjects, with an overall 72.9% reduction in headaches. In the second study, both groups improved with treatment; however the CT group was more likely to experience significant headache relief (72.7%) than the AC group (28.6%, chi 2(1) = 4.97, p < .03). Significant improvement was maintained at a 6-month follow-up for over 50% of patients. It is concluded that the combined nonpharmacological treatment was more effective than an attention control in reducing headaches during pregnancy. This treatment was effective regardless of predisposing variables.

Source text 2 [PMID: 2401622]:
This study investigates the use of biofeedback, relaxation and psychotherapy on five patients with severe, vascular headaches that occurred during the course of pregnancy. The subjects received between four and twelve sessions of treatment overall. The subjects all showed a marked reduction or complete cessation of headaches during treatment, the term of pregnancy, and during a follow-up evaluation months after the birth of the child. Possible alternate explanations for improvement are discussed along with the study's limitations. This preliminary investigation strongly suggests that psychological treatment may be a particularly useful intervention for management of headaches that occur in pregnant women.

Figure 3.12: An example of multi-document summarisation from our corpus, showing the question, the bottom-line summary and two source abstracts.

Question.
What are the best therapies for acute migraine in pregnancy?

SORT grade:
C – based on poor-quality cohort and Case Control Studies. [*PubMed IDs: 8600478, 2401622*]

Source text 1 [PMID: 8600478]:
Concerns about the effects of maternal medications on the growing baby limit the use of medication treatment for benign conditions, such as recurring headaches, during pregnancy and lactation. Nonpharmacological therapies hold particular promise for pregnant women due to the limited medication options. No controlled studies, however, have reported on the efficacy of nonpharmacological treatments for pregnant women. The first study evaluated the effectiveness of a combined nonpharmacological treatment (CT) consisting of relaxation, skin-warming biofeedback, and physical therapy for pregnant women with chronic headaches. In a second study, the CT protocol was compared with an attention control (AC) that received headache education and skin-cooling biofeedback. The first study resulted in significant symptom improvement in 79% of subjects, with an overall 72.9% reduction in headaches. In the second study, both groups improved with treatment; however the CT group was more likely to experience significant headache relief (72.7%) than the AC group (28.6%, chi 2(1) = 4.97, p < .03). Significant improvement was maintained at a 6-month follow-up for over 50% of patients. It is concluded that the combined nonpharmacological treatment was more effective than an attention control in reducing headaches during pregnancy. This treatment was effective regardless of predisposing variables.

Source text 2 [PMID: 2401622]:
This study investigates the use of biofeedback, relaxation and psychotherapy on five patients with severe, vascular headaches that occurred during the course of pregnancy. The subjects received between four and twelve sessions of treatment overall. The subjects all showed a marked reduction or complete cessation of headaches during treatment, the term of pregnancy, and during a follow-up evaluation months after the birth of the child. Possible alternate explanations for improvement are discussed along with the study's limitations. This preliminary investigation strongly suggests that psychological treatment may be a particularly useful intervention for management of headaches that occur in pregnant women.

Figure 3.13: An example of the use of our corpus for automatic grading of evidence.

bottom-line recommendations into three subtasks:

1. Automatic grading of evidence.
2. Single-document, query-focused summarisation.
3. Multi-document, query-focused summarisation.

Automatic grading of evidence is an important subtask of the evidence-based answer generation process. It requires a careful analysis of the evidence relative to the query. This evidence consists of source documents presenting details about relevant studies which can be used to answer the clinical query. Therefore, we incorporate this task in our automatic summarisation research. In our approach, we model this task independently of the summarisation task. In terms of inputs and outputs, the task can be defined as follows:

> **Inputs:** The query (q_i) and the set of source abstracts (D_{ij}) associated with the bottom-line answer (a_{ij}).
> **Output:** A prediction of the quality grade (g_{ij}) for the evidence associated with the bottom-line answer (a_{ij}).

We describe our efforts to automatically grade the quality of medical evidence in Chapter 4.

We model the task of generating single-document summaries as a sentence level extractive summarisation task. The intent is to generate abstracts that most closely resemble the human-authored detailed justifications. In terms of inputs and outputs, the task can be defined as follows:

> **Inputs:** The query (q_i) and a source abstract (d_{ijk}) that is cited in the detailed justification (l_{ijk}) associated with the question.
> **Output:** A summary (s_{ijk}) of the abstract (d_{ijk}) with respect to the query that most closely resembles the human-authored detailed justification (l_{ijk}).

Chapter 5 of this thesis focuses on the generation of single-document summaries.

In Chapter 6, we address the problem of multi-document text summarisation. In particular, we study the possibility of utilising the automatically generated single-document summaries, rather than full abstracts, as source texts for this task. We investigate possible approaches by which information from multiple documents can be fused to generate a bottom-line answer. In terms of inputs and outputs, this task can be defined as follows:

Inputs: A query (q_i) and a set of single-document summaries ($\{s_{ij0}...s_{ijo}\}$) associated with that query.
Outputs: A prediction of the bottom-line recommendations ($\{a_{i0}...a_{in}\}$) associated with q_i.

Details about the models and the approaches we apply are provided in the associated chapters. For the purposes of this research, we address these three problems in isolation from each other and design evaluation approaches for each of these subtasks. Where applicable, we provide the analysis details and results to explain the validity of our models. Figure 3.14 shows a high-level view of our summarisation model. The figure depicts that the bottom-line recommendation generation step (described in Chapter 6) is dependent on the single-document summarisation task (described in Chapter 5). The task of quality appraisal or evidence-based grade recommendation (Chapter 4) can be assumed to occur simultaneously. This is because the evidence quality appraisal step attempts to determine the quality of a collection of studies from which the evidence is obtained and it relies on the information provided by the source abstracts. In other words, the grade recommendation component and the summarisation component of our model receive the same inputs. The processing they perform, however, are different because of the distinct intents of the two components.

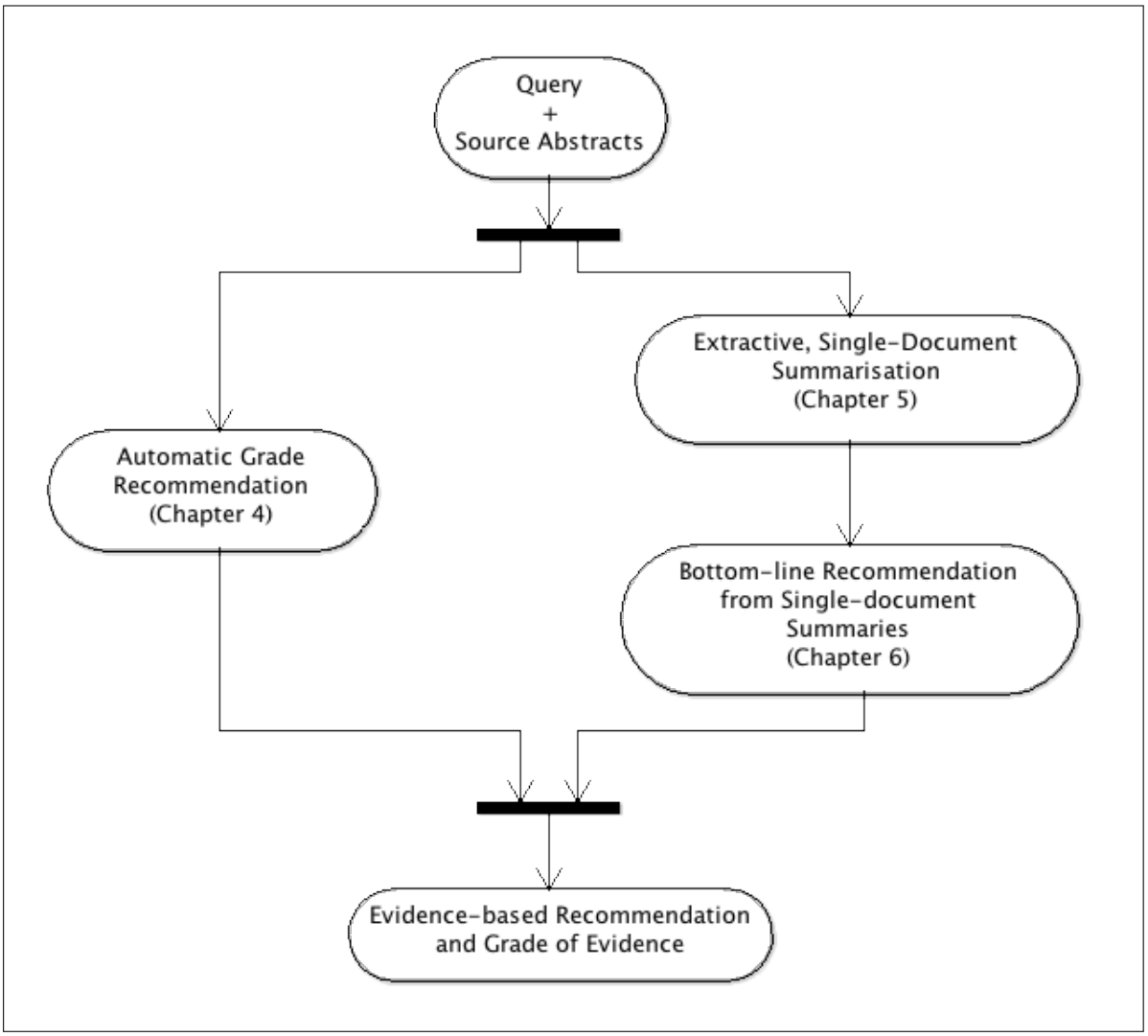


Figure 3.14: The overall summarisation model.

4 Automatic Appraisal of Clinical Evidence

4.1 Introduction

In Chapter 1, we mentioned that our model for the generation of evidence-based recommendations involves two major tasks: summarisation of evidence and appraising the quality of evidence. Appraisal of the quality of evidence is an important task in the process of evidence-based answer generation, and its purpose is to indicate the reliability of the recommendations that are made based on the available evidence. In Chapter 2, we have briefly discussed the task of clinical evidence appraisal and the grading of the quality of evidence in the context of evidence-based medicine. The detailed review of evidence-based medicine practice, provided in that chapter, established the importance of the task of evidence appraisal, and illustrated the strong motivation to automate the process in order to save time. Evidence-based medicine requires the identification and use of the best available evidence, and the grades indicate the qualities of the best available evidences. The quality of the *best* available evidence, however, differs depending on the topic. This is because the volume of research performed on distinct topics can vary. Topics that have received more research attention in the past are likely to contain better quality evidence, compared to topics that have received little. These grades also give an indication of the maturity of research on specific medical topics, suggesting whether further research is needed for the topic. Currently, quality grading is a manual process, requiring the manual analysis of various factors that influence the qualities of medical evidence. The process is time-consuming, and adds to the time requirements of the practice. Thus, the automation of this process is desirable.

In this chapter, we describe our efforts to automate the process of grading the quality of clinical evidence on a chosen scale. Our scale of choice is the Strength of Recommendation Taxonomy (SORT) [Ebell et al., 2004]. From a high-level perspective, our approach relies on identifying key factors that influence evidence grades, and then utilising those factors to estimate the evidence grades. We analyse features that are suggested to be useful in existing literature, and features that

are specifically mentioned to be useful in SORT. In terms of input and output, we formulate the task of evidence quality grading as follows:

Input-1: A clinical query
Input-2: Text from multiple source articles
Output: A SORT grade indicating the quality of the evidence contained in *Input-2*

Despite the importance and necessity of grading the quality of evidence in evidence-based medicine practice, the possible automation of this process has not received much research interest. There has been some research on assessing the qualities of medical publications. However, prior to our work, no work focused on grading against a specialised grading scale such as SORT. Some related research work has recently incorporated the task of evidence appraisal into the summary generation process. For example, Lin and Demner-Fushman [2007] assign evidence grades to their query-focused summaries, but their approach does not take into account a target scale for the grade generation/evaluation process. Our work is the first of its kind in this area, and therefore, we compare the performance of our system against simple baselines and human agreement. The availability of a specialised corpus allows us to analyse the relationships between various factors associated with clinical evidence, and the extent to which the final grades are influenced by these factors. Furthermore, the availability of manual, bias-free annotated data enables us to study supervised classification approaches to recommend evidence grades automatically.

The rest of the chapter is organised as follows. In Section 4.2, we discuss in detail the task of grading the qualities of evidence, and we describe the SORT grading system. In Section 4.3, we present the results of our initial experiments on analysing the factors that influence evidence grades based on the SORT criteria. In Section 4.4, we discuss our experiments of information extraction from medical publications and the final grading of evidence. In Section 4.5, we present a comparison of the performance of our model and system to the grades given by human experts. We conclude with the contributions and findings of this chapter in Section 4.6.

4.2 Grading the Quality of Evidence and the Strength of Recommendation Taxonomy

In our review of the literature for evidence-based medicine practice in Chapter 2, we explained that the key idea behind this practice is to identify the best available clinical evidence associated with a clinical query before making a decision. This requires exploring the clinical research literature relevant to a topic and identifying the conclusions obtained. The amount of research literature available varies with respect to the topics. While some topics may have received a large amount of structured clinical research, others may not have been explored thoroughly, and the

best evidence regarding those topics may still be insufficient or incomplete. In other words, the reliability of the evidence associated with different topics may vary depending on the amount of research the topics have received. Also, sometimes, findings from different studies are not consistent, making the evidence unreliable. When making evidence-based recommendations, practitioners have to take these and other factors into account, and assess the reliability of the extracted evidence. Thus, when extracting evidence from medical publications regarding a topic, practitioners also have to appraise the quality of the evidence associated with the topic. This quality of evidence depends on all published research relevant to a specific topic, and, therefore, may be based on a single study or a collection of studies. Typically, however, the quality of evidence is based on a body of evidence consisting of more than one study. This is also depicted by the distributions of references for bottom-line summaries shown in Figure 3.9 of Chapter 3 (i.e., a very small proportion of bottom-line summaries have only one associated source document).

Due to the importance of identifying and specifying grades indicating the qualities of evidences, standardised grading scales have been proposed in the literature. These grading scales generally specify various criteria based on which the grade of an evidence is presented on a discrete scale. In our corpus, the grades are presented using a scale known as the Strength of Recommendation Taxonomy (SORT). SORT was first proposed in 2004 through a collaborative effort by the editors of multiple family medicine journals with the purpose of providing authors and readers of family medicine journals with a simple, user-friendly system for the grading of evidence. SORT was designed to provide a uniform recommendation-rating system that could be applied throughout the medicine literature. It is simple and straightforward, and, therefore, easy for practitioners to use during everyday practice. This taxonomy uses only three ratings – A (strong), B (moderate) and C (weak) – to specify the strength of recommendation of a body of evidence. Explanations of these grades have been provided in Chapter 2. We review them in more detail here, before we dive deeper into our experimental details:

- Grade **A** – This reflects a recommendation based on *consistent* and *good-quality*, *patient-oriented* evidence. A collection of studies is considered to be *consistent* if the results of most of the studies are found to be similar, have explainable differences or if the recommendation is supported by high quality and up-to-date Systematic Reviews or Meta-Analyses; a *good-quality* evidence consists of high quality Systematic Reviews, Meta-Analyses, Cohort Studies with good followup, or high quality Randomised Controlled Trials; *patient-oriented* evidence measures outcomes that matter to patients: morbidity, mortality, symptom improvement, cost reduction, and quality of life.

- Grade **B** – This reflects a recommendation based on *inconsistent*, *limited-quality*, *patient-oriented* evidence. A collection of studies is considered *inconsistent* if there is considerable variation among study findings and inexplicable differences among them, or if the evidence

does not directly favour the recommendation. Limited-quality evidence consists of lower-quality Meta-Analyses, Cohort Studies and Clinical Trials, Case Control Studies and so on.

- Grade **C** – This reflects a recommendation based on consensus, usual practice, opinion, *disease-oriented* evidence, or Case Series for studies of diagnosis, treatment, prevention or screening. *Disease-oriented* evidence measures intermediate, physiologic, or surrogate end points that may or may not reflect improvements in patient outcomes (*e.g.*, blood pressure, blood chemistry, etc).

Further details about the grades can be found in the paper by Ebell et al. [2004]. The paper also provides specific guidelines that must be followed when grading evidence on this scale. These guidelines present elaborate procedures customised to different types of studies such as *diagnosis*, *treatment*, *prevention*, *screening*, and *prognosis*. Furthermore, the authors provide two algorithms for grading qualities of (i) bodies of evidence, and (ii) individual studies. These algorithms are shown in Figures 4.1 and 4.2 respectively. In the following sections, we address the possibility of automatically generating evidence grades using lexical features and meta-data from the documents in our corpus.

4.3 Factors Influencing Evidence Grades

We commence our work on automatic grading of evidence by performing an analysis of the factors that influence the quality of evidence. We focus on features from individual source documents and apply them to predict the associated quality grades. In Chapter 3, we explained how our corpus is suited for this task, and we utilise various information encoded in the corpus to perform this analysis. In particular, we use the grades encoded in our corpus as the gold standard, and we analyse the extent to which these grades can be automatically predicted using various other information also encoded in the corpus.

Ebell et al. [2004] mention a number of factors that are relevant for grading evidence based on the SOR taxonomy. These factors include: quality of evidence of the individual studies, types of evidence presented in the studies (i.e., patient- *vs.* disease-oriented), and consistency of outcomes presented in different studies. Besides analysing the effects of some of these factors, we also experiment with other factors that have been shown/suggested to be useful in the literature. Such factors include: publication venues of the individual documents (*e.g.*, journal names), and publication dates.

Research work related to this task has focused mostly on automatic quality assessment of medical publications for purposes such as retrieval and post-retrieval re-ranking, where approaches

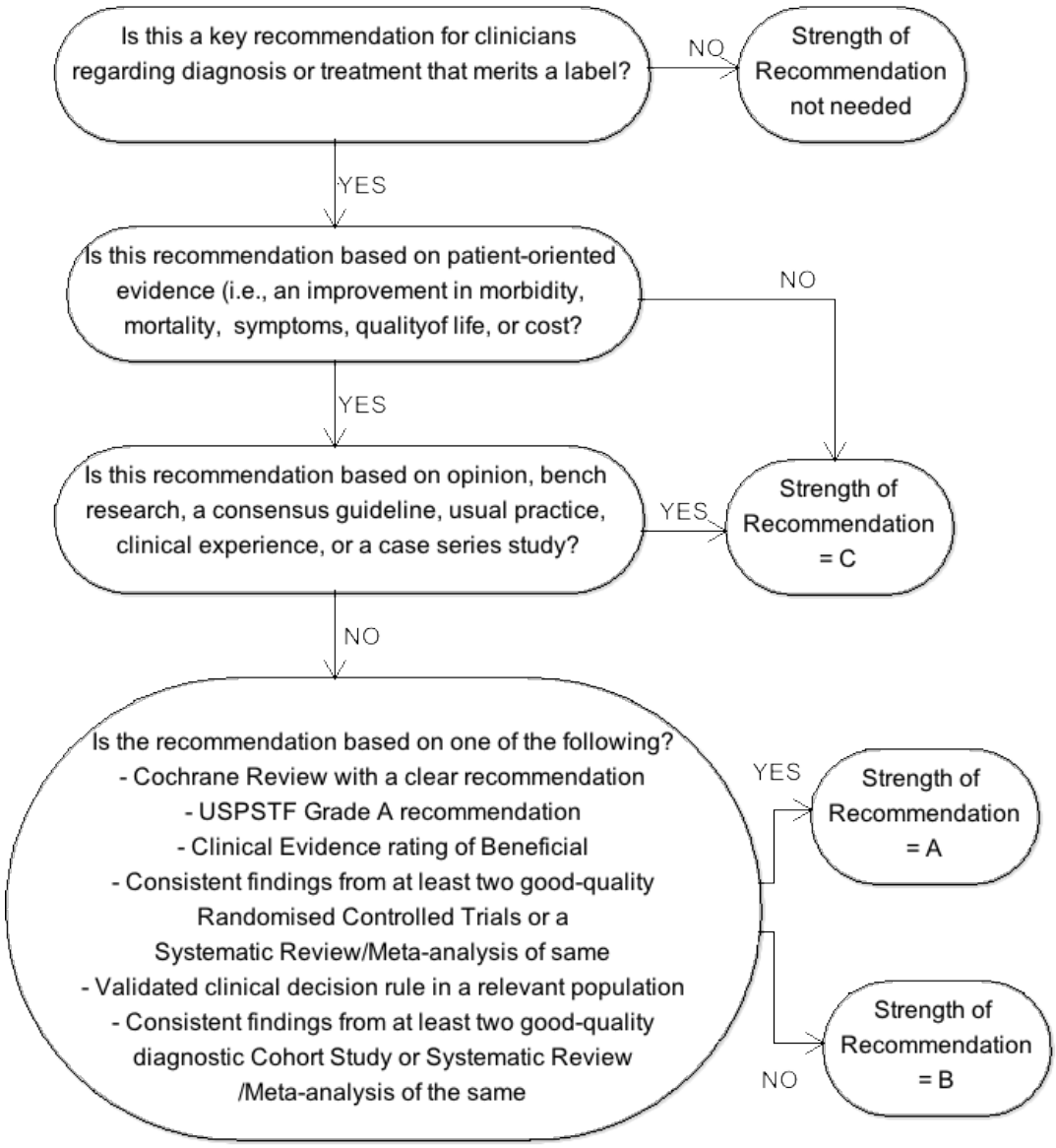


Figure 4.1: Algorithm for determining the strength of a recommendation based on a body of evidence. Source: Ebell et al. [2004].

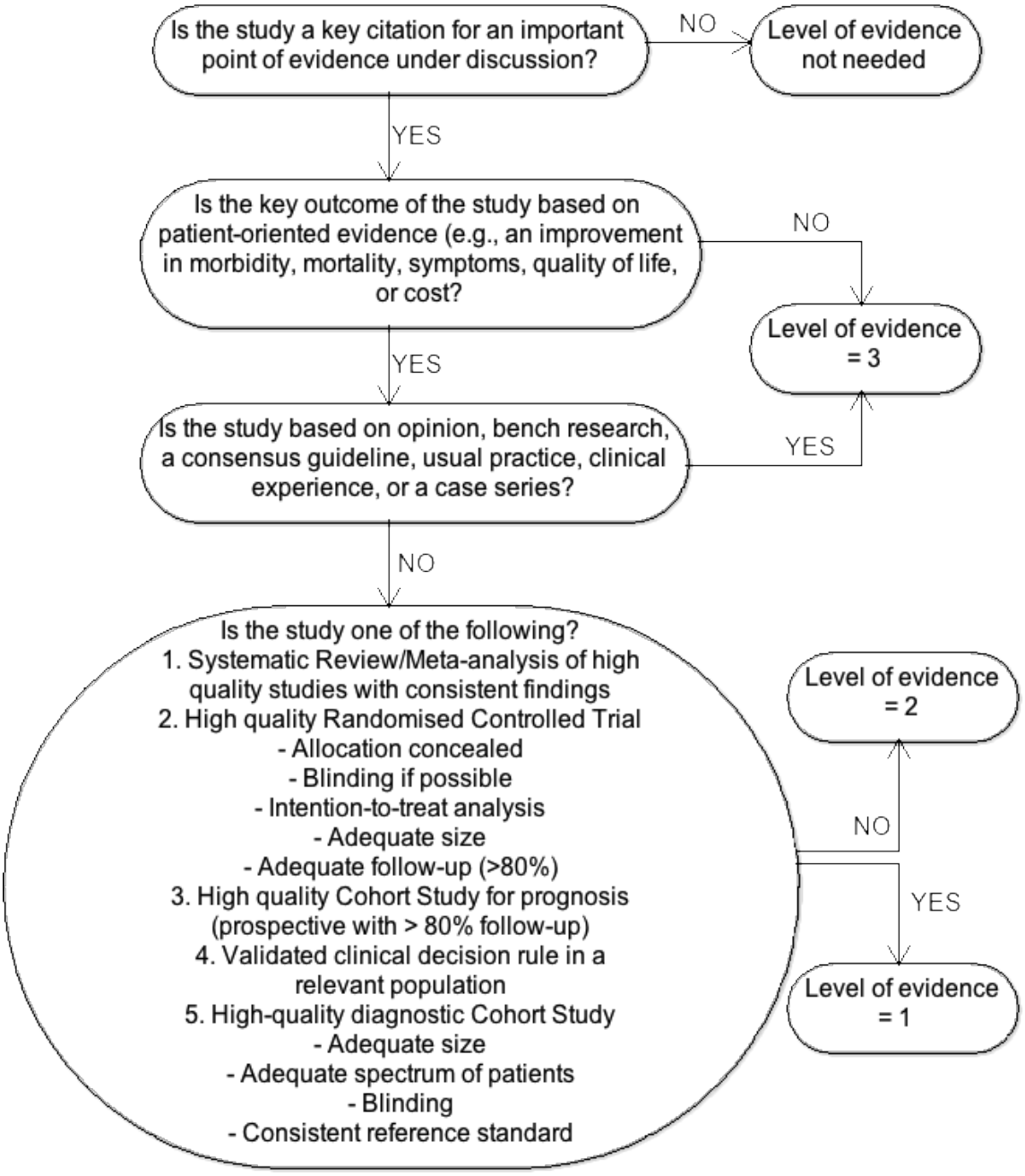


Figure 4.2: Algorithm for determining level of evidence for an individual study. Source: Ebell et al. [2004].

based on word co-occurrences [Goetz and von der Lieth, 2005] and bibliometrics [Plikus et al., 2006] have been proposed for improving the retrieval of medical documents. These approaches, however, do not integrate evidence-based recommendations for appraisal. Tang et al. [2009] propose a post-retrieval re-ranking approach that attempts to re-rank results returned by a search engine. The results may or may not be published research work. Their approach is only tested in a specific sub-domain (i.e., Depression) of the medical domain. Kilicoglu et al. [2009] focus on identifying high quality medical articles and build on the work by Aphinyanaphongs et al. [2005]. They apply machine learning and obtain 73.7% precision and 61.5% recall. These approaches and related research generally model the problem of quality assessment as a binary classification task, where each article may either be of *good* or *bad* quality. Also, the approaches are suitable for single documents only. Our research work has two primary differences with existing research on automatic quality assessment: (i) we use a more standardised and specialised scale, with the intent of automatically recommending evidence-based grades; (ii) our approach is for *bodies* of evidence, which may be single documents or multiple documents on the same topic.

Ebell et al. [2004] suggest that the publication types of medical articles are good indicators of their qualities. Literature in the medical domain consists of a large number of publication types of varying qualities[1]. For example, a Randomised Controlled Trial is of much higher quality than a case study of a single patient. Evidence obtained from the former is thus more reliable. Greenhalgh [2006] mentions some other factors that influence the grade of an evidence, such as the number of subjects included in a study and the mechanism by which subjects are allocated (*e.g.*, randomisation/no randomisation), but the latter is generally indicated by the publication type (*e.g.*, Randomised Controlled Trial) of the article. Lin and Demner-Fushman [2007] also acknowledge the importance of publication types in determining the quality of clinical evidence. They use a working definition of the 'strength of evidence' as a sum of the scores given to journal types, publication types and publication years of individual publications. Their scores are used for citation ranking, not evidence grading, and therefore their results cannot be compared to ours. However, their research does suggest that the journal names and publication years have an influence on the qualities of individual publications, which in turn may influence the grade of evidence obtained from them.

4.3.1 Analytical Methods

In this analysis, we model the problem of evidence grading as a text classification problem, where the grades are considered to be classes for source articles (evidence) associated with them. This formulation enables us to assess the validity of two propositions:

[1] A list of publication types used by the U.S. National Library of Medicine can be found at http://www.nlm.nih.gov/mesh/pubtypes2006.html. Accessed on 26th May, 2014. This list is not exhaustive.

1. Evidence grading may be modelled as a text classification problem.
2. Features may be extracted from source documents to automatically predict the qualities of the evidences they present.

We perform this analysis using the grade annotations in our corpus. We use a sample consisting of parts of the records that have the evidence grades A, B, or C associated with them. We ignore recommendations that either have no grades associated with them or have non-standard SOR grades, such as the grade D[2]. For each recommendation having one of the abovementioned grades, we collect information associated with them from our corpus.

Our data set for these experiments consists of 1,132 evidence-based recommendations generated from a total of 2,713 medical documents. Of the 1,132 recommendations, 330 are of grade A, 511 of grade B and 291 of grade C. Based on the guidelines provided by Ebell et al. [2004], and suggestions from the related literature, we focus on analysing the effects of the following features in determining the evidence grades:

1. The ***publication types*** of the individual publications associated with the recommendations. In our corpus, the publication types associated with bottom-line recommendations are generally mentioned (as is the case with the articles from the JFP from which the corpus has been collected)[3]. For the cases where the publication types are not mentioned, we attempt to manually identify the publication types of the documents associated with the recommendations and use that information in our analysis. The publication types used by the JFP do not fully intersect with the publication types used by the Medline database. We discuss this in detail later in this chapter.
2. The ***publication years*** of the associated documents. Related literature suggests that recent publications are more relevant/reliable than older publications [Lin and Demner-Fushman, 2007]. To assess if that really is true, we incorporate this information in our analysis. We obtain publication years of the documents from the PubMed XML files associated with the bottom-line recommendations and use them as numeric features.
3. The ***publication venues***. We include this feature to assess if studies published in high-impact journals give better quality evidence compared to studies published in venues of lower impact. Past research suggests that the qualities of individual studies may depend on the publication venues, and, therefore, this information is likely to affect the overall evidence grade [Lin and Demner-Fushman, 2007]. Information about the publication venues are also encoded in the PubMed XML files present in our corpus.

[2] Although D is not a standard grade in SORT, some JFP contributors used this grade to indicate very low or unclear levels of evidence.

[3] An example of this can be found at: http://www.jfponline.com/pages.asp?aid=8103 (Accessed on: May 13, 2014)

4. The **titles** of the articles. The titles of the articles present useful information such as the topics of the studies, and, often, other information such as disorders, interventions and so on. The titles are also obtained from the PubMed XML files.

Note that, for this preliminary analysis, we refrain from using text from the actual abstracts. This is because using word n-grams from the abstracts introduces a large number of features, making the analysis of other features difficult. We utilise lexical information from the abstracts in the work described later in this chapter.

Due to the large number of possible publication types that medical articles can have, we group together publication types having low frequency and similar quality levels, since it is not possible to accommodate all publication types. This grouping is performed manually. Our final set consists of 11 groups of known publication types, each having a different quality level, and 1 group of unknown types, as shown in Figure 4.3. Based on our collected data, we consider 45.1% — the accuracy when all instances are classified as B (the majority class) — as the baseline for our experiments.

We study the distribution of publication types over the SOR grades. This distribution is shown in Figure 4.3. In the Figure, *Other Study* refers to low frequency studies (*e.g.*, *Observational Study*), *Other Clinical Trial* refers to all clinical trials other than *Randomised Controlled Trials*, and *Unknown* refers to articles with unidentified publication types. A clear pattern in the distribution of publication types over SORs can be seen. For SOR A, evidence primarily comes from Randomised Controlled Trials, Systematic Reviews and Meta-Analyses, and the numbers drop significantly for other publication types. For SOR C evidence, most of the evidence comes from publications presenting Expert Opinion, Case Series/Reports, and Consensus Guidelines. The distribution for SOR B has the largest spread with Cohort Studies having the highest frequency. The distributions suggest that the publication types play an important role in determining the SOR.

4.3.2 SOR Prediction from Publication Types

To test the extent to which SORs can be predicted from the publication types, we perform basic experimentation using supervised machine learning. We model the grading of evidence as a single-label text classification problem with three classes. Since the SOR grades are provided in our corpus, we apply various supervised machine learning algorithms for the task. Due to the mentioned importance of medical publication types [Ebell et al., 2004], we first experiment using only the publication types of the articles as features. Each instance in our model represents an evidence-based answer composed of the SOR class and a vector containing the counts of each of the 12 publication types shown in Figure 4.3. Therefore, the task of each classifier is to predict

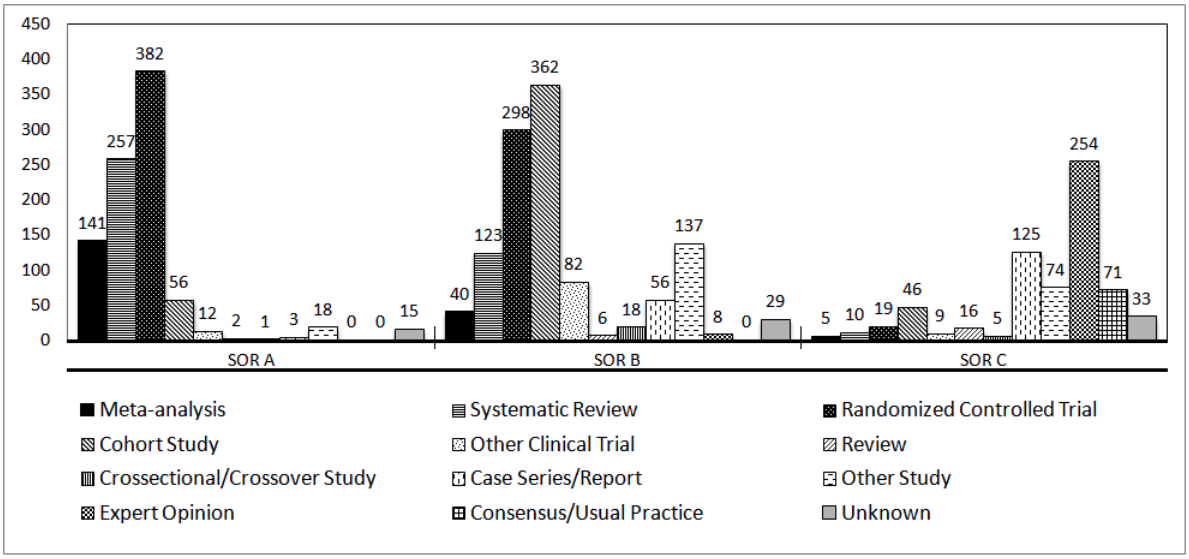


Figure 4.3: Distribution of publication types across SORs.

the class of each instance given the associated publication type information.

We use two-thirds of our data for training and the remaining as held-out test data. For both sets, we keep the proportions of instances belonging to the three classes the same as their proportions in the whole data set. We perform our experimentation using the software package Weka[4]. Weka provides implementations of a range of classifiers organised into generic groups, and in our preliminary analysis, we experiment on our training data with multiple classifiers belonging to each generic group. We choose five classifiers that produce good results on our training data and have also been shown to produce good results on similar problems in the past. The five chosen classifiers are (the names used in Weka shown in parenthesis): Bayes Net, Support Vector Machines (SMO), K-Nearest Neighbour (IBk), Multinomial Logistic Regression (Logistic) [Le Cessie and Van Houwelingen, 1992][5], and C4.5 Decision Tree (J48) [Quinlan, 1993].

For specific classifiers, we perform simple parameter tuning and choose parameter values that produce the best results for stratified 10-fold cross validations on the training set. For the Bayes Net classifier, we use the K2 search algorithm [Cooper and Herskovits, 1992] for local score metrics and the simple estimator for estimating conditional probability tables. For the SMO algorithm, we use John Platt's [1998] sequential minimal optimisation algorithm and solve our multi-class problem using pairwise (1-*vs.*-1) classification. We use an RBF kernel for the SMO classifier, normalising all attributes and using a grid search to find good values for the parameters

[4]http://www.cs.waikato.ac.nz/ml/weka/. Accessed on 26th May, 2014.

[5]The Weka implementation of this algorithm is slightly different from the original implementation. Details can be found at: http://www.java2s.com/Open-Source/Java-Document/Science/weka/weka/classifiers/functions/Logistic.java.htm. Accessed on 26th May, 2014.

Classifier	Accuracy (%)	95% CI	Parameters
Bayes Net	66.578	61.6-71.3	$K2, SimpleEstimator$
SMO	68.449	63.5-73.1	$\gamma = 1.0, C = 2^7$
K-Nearest Neighbour	68.717	63.8-73.4	$K = 7$
Logistic Regression	67.380	62.4-72.1	..
C4.5	68.182	63.2-72.9	$confidenceFactor = 2^{-1}$

Table 4.1: Accuracies, 95% confidence intervals and specific parameter values for various classifiers, using only publication types as features.

γ and C. To find the best value of K for the K-Nearest Neighbour algorithm, we search through all odd values of K from 1 to 101. For the C4.5 Decision Tree classifier, we search between 2^{-5} and 2^{-1} to find the best value for the confidence factor parameter.

4.3.3 SOR Prediction from other Factors

In addition to the use of publication types as features, we use the journal titles, publication years, and article titles as features. While journal titles and publication years have been used for assessing quality in the past [Lin and Demner-Fushman, 2007], article titles have not been used. However, we suspect that titles may help to identify the qualities of individual publications, since they sometimes provide useful information about how the studies are carried out (*e.g.*, 'A Double-blind, Placebo-controlled Trial'). In our model, we represent the titles and journal names using uni- and bi-grams. Prior to generating the n-grams, we process the titles by removing stop words, lowercasing the words, stemming the remaining words using the Porter stemmer [Porter, 1980] and removing words occurring less than five times across the whole data set. We repeat the experimental procedures mentioned above with various combinations of these feature sets.

4.3.4 Analysis Results

Using only publication types as a feature set, we obtain classification accuracies of approximately 66 - 69% (over 20% improvement over the baseline) with various classifiers on our held-out test set. Table 4.1 shows the accuracies of the five abovementioned classifiers along with 95% confidence intervals[6] for the accuracies and important parameter values for specific classifiers. There is no statistically significant difference between the accuracies of the different classifiers, meaning that the most important role to determine classification accuracy is played by the features used, not the type of classifier.

A manual content analysis of the incorrect classifications reveals that there were few errors

[6]Calculated using the package R's binom.test function (http://www.r-project.org. . Accessed on 26th May, 2014).

between A and C, which is exactly what is expected based on their very different distributions of publication types. The most common errors are between SOR A and B, and SOR C classified as B. Our manual analysis reveals that errors are caused primarily by factors such as sizes of studies, consistency and types of outcomes, which our classifiers do not take into account. For example, an essential condition for an evidence to be of grade A or B is the presence of patient-oriented outcome, irrespective of the type of study. At the same time, for certain types of publications, such as Cohort Studies, the sizes of the studies significantly influence their quality. Unaware of this information, our classifiers classify all evidences obtained primarily from Cohort Studies as grade B. Furthermore, evidence obtained primarily from Meta-Analyses and Systematic Reviews are graded as A, irrespective of the consistency or types of outcomes presented in the studies. Lastly, due to the smaller number of C grade instances, the classifiers often tag C grade evidence as B.

Features	**Accuracy (%)**	**95% CI**	**Classifier**
Journal, Pub. Year, Title and Pub. Type	63.636	58.5-68.5	C4.5
Pub. Type and Pub. Year	66.578	61.6-71.3	C4.5
Pub. Type and Title	67.380	62.4-72.1	C4.5
Pub. Type and Journal	63.904	58.8-68.8	C4.5
Journal, Pub. Year and Title	50.802	45.6-56.0	SMO
Journal and Pub. Year	46.257	41.1-51.5	SMO
Title only	51.070	45.9-56.2	SMO
Pub. Year only	47.594	42.4-52.8	Bayes Net
Journal only	47.326	42.2-52.5	Bayes Net

Table 4.2: Accuracies, 95% confidence intervals, and best performing classifiers for various feature sets.

Our experiments suggest that adding factors such as journal names, publication years and article titles to the publication types do not significantly influence the SORs. Table 4.2 shows the highest accuracies obtained using various combinations of feature sets, the 95% confidence intervals and the classifiers with the best results. From the table it is evident that the absence of publication types as a feature set causes significant drops in accuracy. Although incorporation of article titles as a feature set produces marginally better accuracies compared to our baseline, no significant improvements are achieved when this feature set is combined with publication types. The other feature sets, alone or in combination with each other, do not give a statistically significant improvement over the baseline.

Our experiments show that significant improvements over the baseline accuracy can be obtained using supervised classification. Based on this preliminary analysis, we conclude that evidence grading may be modelled as a text classification problem, which can potentially be solved using supervised machine learning. In the analysis, only the publication type feature appeared to be

useful, while the other features did not provide any significant improvements. Also, we only use a small number of manually specified features for this task. In a fully automatic evidence grading task, more features can be extracted from the texts of the abstracts associated with the evidence-based answers. Since the publication types appear to be most useful, we address the problem of automatically detecting the publication types of medical articles and applying that information for classification. We then combine the automatically extracted publication type information with other lexical features to automatically grade qualities of evidence.

4.4 Automatic Grading of Evidence

In our approach for fully automatic appraisal of medical evidence, we first address the issue of automatic feature extraction. Our preliminary analysis, described above, shows that information about the publication types of medical articles play a vital part in determining the evidence grades automatically. We, therefore, attempt to devise a strategy to automatically determine the publication types using lexical information from the abstracts of the articles, and the meta-data accompanying the texts in the PubMed abstracts. Following that, we attempt to perform fully automatic grading of evidence using a combination of various automatically extracted features. We detail these approaches in this section.

4.4.1 Identifying Publication Types of Medical Articles

The XML files representing source abstracts in our corpus have, in addition to the titles and the texts, some associated meta-data. These meta-data are added to the articles when indexing them on PubMed. Each abstract has at least one *PublicationType* tag, the intent of which is to state the publication type of the article. However, not all publication types that we require are present in the PubMed articles. For example, PubMed does not have a specialised tag for Systematic Reviews. The *Review* tag is used to represent both Systematic Reviews and non-systematic Reviews. In many cases, simply the default *Journal Article* tag is used for the articles. There is also no specialised tag for Cohort Studies, although a large number of articles in our data set belong to that broad category. Because of these reasons, we can not rely fully on the meta-data associated with the articles to identify the publication types.

We apply a simple, rule-based approach to automatically identify several important publication types. We combine the publication types identified by this approach along with the publication type information provided as meta-data. In our approach to identify the publication types, we only use texts from the article titles and abstracts. Abstracts, and often titles, of medical articles contain information about the types of studies and therefore provide evidence of their publication types. Figure 4.4 presents three examples of evidence of Randomised Controlled Trials. The first

example shows evidence from title. The second and third are examples of sentences from article abstracts that contain evidence. The important contents of the sentences are shown in bold.

Evidence of publication type in title:

*A **randomised controlled trial** of self-help interventions in patients with a primary care diagnosis of irritable bowel syndrome.*

Evidence of publication type in abstract:

***Prospective randomised controlled trial** of low risk women admitted in spontaneous labour, with intact membranes.*

*In this study, 200 participants who met the diagnostic criteria for cervicogenic headache were **randomised into four groups**: manipulative therapy group, exercise therapy group, combined therapy group, and a control group.*

Figure 4.4: Examples of evidence of publication type in title and abstract texts. The first example shows how the title can provide evidence of publication type. The second and the third examples show how abstract sentences can provide evidence about the publication type.

Our approach relies on regular expressions to identify relevant patterns (evidence) from titles and abstracts. We provide details of our approach in the following subsections. In addition to developing regular expressions for publication types that do not have specific tags in PubMed, we implemented some expressions for publication types that have specific tags. This is for two reasons:

i often multiple publication types are assigned to a single article, and our expressions were aimed at detecting the most relevant tag for an article; and

ii in some cases, the tags assigned on PubMed may not be consistent, and applying our regular-expression based classifier, the publication type can be detected with more accuracy for certain publication types.

An example of (i) may happen when an article is tagged as a *Randomised Controlled Trial*, *Controlled Clinical Trial*, and *Clinical Trial* in the PubMed meta-data. Using our rule-based approach, if the article is found to be a Randomised Controlled Trial, only that tag is kept and the others are discarded. Details of this is provided later in this section when we discuss our features for the machine learning classifiers in detail. As an example of (ii), we found a number of cases

```
Evidence of randomisation for Randomised Controlled Trials:

random.*alloc
random.*chose
chose.*random
random.*assign
assign.*random
random.*appli
appli.*random
desig:.*random
animal.random
random.*animal
patien.*random
subjec.*random
randomi[sz].*group
parallel[\W]*group
group.*random
random.*doub.*blind
random.*open[\W]*label
randomi[sz]e.*trial
doubl.*blin

Evidence of no or unacceptable randomisation:
coin\W*flip
non\W*random
odd\W*even
uncontrol\W*stud
```

Figure 4.5: Sample patterns used for detecting Randomised Controlled Trials. Patterns used for detecting unacceptable randomisation techniques are also shown.

where an article was tagged as *Multi–Centre Study* only, while it was actually a Multi-Centre Randomised Controlled Trial. The intent of the rule-based approach, in this case, is to tag the article as a Randomised Controlled Trial since that is the most relevant for our work.

We develop regular expressions to classify articles by manually studying the titles and abstracts of articles belonging to several publication types. We collect our development set from a mixture of sources. For articles which have associated *PublicationType* tags in PubMed (*e.g.*, Randomised Controlled Trials and Meta-Analyses), we retrieved about two hundred of each type. We study each article individually, identify the evidence of publication type and develop patterns to pick up the evidence. During the development of the rules, we use an incremental approach similar to the Ripple Down Rules [Compton and Jansen, 1988] philosophy – after adding a new regular

```
Evidence of Meta-Analyses (in title or text) of article:
'meta[-]analys'

Evidence of Systematic Reviews:
sytemat.*rev
cochr.*medlin
search.*cochr
search.*embas
search.*medlin.*datab
cinahl.*search
literat.*embas
medic.*liter.*rand

Evidence of Cohort Studies:
retrospect.*stud
foll.*prospect
cohort.*stud
retrospect.analys
patien.*foll.*year
foll.*cohort
retrospect.*trial
prospect.*analys

Evidence of Practice Guidelines:
guidel.*diag.*treat
clinic.*guidel

Evidence of Consensus Development Conferences:
consens.*confer
consens.*statem
consens.*devel
reac.*consens

Evidence of Review (non-systematic):
postmar.*survei.*surv
retrospec.*char.*rev
```

Figure 4.6: Sample patterns used for detecting specific publication types.

expression we test its effect on our development set and add more expressions based on the articles that are not correctly identified. We utilise this strategy for Meta-Analyses, Randomised Controlled Trials, Consensus Development Conferences, Practice Guidelines, and Reviews. For example, in the case of Randomised Controlled Trials, we primarily develop expressions to detect evidence of randomisation in the abstracts. Once evidence of randomisation is found, we also develop expressions (from false positives) to detect evidence(s) of unacceptable randomisation[7]. Some of the expressions we use to identify Randomised Controlled Trials and unacceptable randomisation techniques are shown in Figure 4.5 (the list is not exhaustive).

For articles without an associated *PublicationType* tag in PubMed (*e.g.*, Systematic Reviews), obtaining a large development set is considerably more difficult. We therefore use a mixture of secondary sources of evidence such as the Journal of Family Practice and the Cochrane Library for obtaining about fifty of each and develop our expressions from that set. In our corpus, the publication types of the cited articles are often given, and we use those annotations as our gold standard. Furthermore, we study search techniques suggested by PubMed[8] for the efficient retrieval of Systematic Reviews and develop expressions based on their suggestions. We also develop expressions based on search keywords and techniques suggested in the literature for obtaining articles of specific publication types [Hunt and McKibbon, 1997, Montori et al., 2005]. Developing rules is easier for publication types of higher qualities (*e.g.*, Systematic Reviews). This is primarily because articles belonging to these publication types often have standard discourse structures, and also, almost invariably, clearly state the type of publication in the abstract or the title. The same is not true for publication types of lower qualities.

We apply this approach for Systematic Reviews and Cohort Studies. We also attempt to apply this technique for Case Series, Case Control Studies, and some other publication types. However, due to the inconsistent nature of the abstracts of articles belonging to these publication types, and the variety of ways in which the abstracts are written, the accuracies achieved by our preliminary experiments are low. Therefore, for such lower quality publication types, we primarily rely on the *PublicationType* tags provided by PubMed. Figure 4.6 shows some of the regular expressions used for detecting Systematic Reviews, Meta-Analyses, and some other publication types (the list is not exhaustive).

We apply a decision list to identify the publication types of articles. Each article is initially assigned an empty tag and passed through a sequence of tests, each responsible for checking for patterns indicating a specific publication type. At any stage of the sequence, both the PubMed *PublicationType* tag and the regular expressions corresponding to a specific publication type

[7]Details about unacceptable randomisation techniques for Randomised Controlled Trials and other publication types can be found at http://www.nlm.nih.gov/mesh/pubtypes2004.html. Accessed on 26th May, 2014.

[8]The techniques can be found at http://www.nlm.nih.gov/bsd/PubMed_subsets/sysreviews_strategy.html. Accessed on 26th May, 2014.

are utilised to check if an article belongs to that category. For the regular expression matching, the title of the article is first checked, and, if no evidence is found, the abstract is checked. If sufficient evidence of a particular publication type is found (with no further evidence of negation), the article is tagged and removed.

The sequence in which the operations are applied is very important as the number of false positives may increase significantly if the sequence is changed. For example, if Systematic Reviews and Meta-Analyses are not removed before searching for Randomised Controlled Trials, many of the former are falsely tagged as the latter. This is because abstracts of Systematic Reviews and Meta-Analyses are often produced by collecting information from multiple Randomised Controlled Trials, and they usually mention the number and types of studies being reviewed/analysed (*e.g. We conducted a systematic review of five double-blind, randomised controlled trials to investigate ...*). Thus, both these publication types generally mention other publication types and must be removed early on in the sequence. The following list elaborates the actions performed at each stage of the sequence:

1. Check for evidence of Meta-Analysis
2. Check for evidence of Systematic Review (regular expressions only)
3. Check for evidence of Review (non-systematic) and Consensus Development Conference
4. Check for evidence of Guideline/Practice Guideline
5. Check for evidence of Randomised Controlled Trial
6. Check for evidence of other forms of clinical trials (meta-data only)
7. Check for evidence of Cohort Studies (regular expressions only)
8. Check for evidence of other forms of studies (*e.g.*, Evaluation Studies, Cross-sectional Studies, Multi-centre Studies, Case Reports, etc.).

In all cases, the meta-data is checked first (if available), then the title and finally the abstract text. While checking the abstract of an article for evidence, each sentence is searched separately. We have attempted other approaches such as searching the whole abstract and using a sliding window. However, we have found sentence level searching to produce the best results primarily because evidence of publication or study type is usually stated or described in a single sentence of an article abstract. Once a pattern match occurs, the entire abstract is searched again to identify patterns that negate the evidence in specific cases (such as unacceptable randomisation techniques in the case of Randomised Controlled Trials), and the article is only tagged if no evidence of negation is found.

To evaluate the performance of our approach, we required a set of test articles that were different from the development set and at the same time completely reliable. To achieve this, we use the articles in our corpus (obtained from JFP) that are explicitly mentioned by the JFP authors to belong to specific publication types. We do not use articles that were used to prepare the development set. Importantly, the chosen articles are not actually written by JFP authors, but are cited by them within JFP articles which provide evidence-based answers to clinical queries. Hence, the chosen articles come from a variety of sources, and this enables us to test our approach on a diverse article collection. As explained in Chapter 3, the article abstracts (along with the title, text, and meta-data) were collected for our corpus from Medline, and we add them to the test set after manually annotating them based on the JFP classifications. Relying on JFP for the test data also allow us to include articles from a wide range of medical topics, thus ensuring that our approach is not topic dependent. To further prevent bias, all articles identified are added to the test set regardless of their structure/content, and the abstracts of the articles are not reviewed during the annotation process. Such a labourious annotation process is necessary due to the lack of substantial reliable annotated data.

In our initial test set, we use a total of 294 articles including 111 Systematic Reviews and Meta-Analyses, 100 Randomised Controlled Trials and 83 articles belonging to a mix of other publication types. Our intent is to evaluate the performance of our approach for these three publication types only (since these are generally associated with the SOR grade A). Including a set of articles belonging to various other publication types is necessary to ensure that our approach does not only correctly tag Systematic Reviews, Meta-Analyses, and Randomised Controlled Trials, but also ignores other types of articles. The recall, precision and F-score values are shown in Table 4.3. Note that in the table, we do not separate between Systematic Reviews and Meta-Analyses because these two publication types are essentially the same (i.e., Meta-Analyses are types of Systematic Reviews). For Systematic Reviews and Meta-Analyses, our approach produces perfect precision but fails to identify one Systematic Review. Our approach tags a total of 97 articles as Randomised Controlled Trials, of which 96 are correctly identified.

Publication Type	**Recall**	**Precision**	**F-Score**
Meta-Analysis and Systematic Review	0.990	1.00	0.995
Randomised Controlled Trial	0.960	0.990	0.975

Table 4.3: Automatic classification results for Systematic Reviews, Meta-Analyses, and Randomised Controlled Trials. Sample size = 294.

In the case of Randomised Controlled Trials, the falsely tagged article is a Review (non-systematic) which mentions '*one randomised, placebo-controlled study*' and is therefore picked up by our rules. As for the four Randomised Controlled Trials that are not identified, none of their

abstracts contain any evidence of randomisation although for one of the Randomised Controlled Trials, there is clear evidence of randomisation in the full article text. In the case of Systematic Reviews and Meta-Analyses, the unpicked article is a Systematic Review in which the abstract does not contain any detail of the study type.

Following this, we perform experiments to evaluate the accuracies for other publication types. The number or articles in each category, however, is lower. We use a total of 78 Cohort Studies, 17 Consensus Development Conferences (CDC), 31 Practice Guidelines (PG), 92 Non-randomised Clinical Trials (Other CT), and 89 other studies (Other). Table 4.4 presents the performance of our approach for these publication types. It can be seen that the F-scores for these publication types are lower than those for the higher quality publication types.

Publication Type	**Recall**	**Precision**	**F-Score**
Cohort	0.81	0.78	0.795
CDC	0.76	0.92	0.832
PG	0.90	0.86	0.883
Other CT	0.84	0.79	0.814
Other	0.78	0.65	0.711

Table 4.4: Automatic classification results for Cohort Studies, Consensus Development Conferences (CDC), Practice Guidelines (PG), Non-randomised Clinical Trials (Other CT), and other publication types (Other). Sample size = 307.

The results indicate that a rule-based approach such as ours is very effective in classifying high quality publication types, such as Systematic Reviews, Meta-Analyses, and Randomised Controlled Trials. The high F-scores can be attributed to the fact that articles belonging to these three publication types are very structured (since there are very specific guidelines that must be followed when writing these articles), and therefore, their titles and abstracts almost invariably contain sufficient evidence of the type of publication, which can be automatically identified. The recall and precision values decrease for other publication types, particularly those of lower qualities. However, for most of these publication types (other than Cohort Studies), the PubMed *PublicationType* tag is generally correctly specified. Thus, with these two information combined, it is possible to detect the abovementioned publication types fairly reliably.

4.4.2 Features and Methods for SOR Classification

In line with our preliminary research work, we model the problem of evidence grading as a supervised classification problem with three classes and attempt to solve it via machine learning algorithms. We use the data from the 2011 ALTA shared task [Mollá and Sarker, 2011]. The task

was based on the problem of automatic grading of evidence, and the data set was prepared from our corpus. The data for the shared task consisted of a set of 'evidences' with the SORT grade for each. Each evidence was represented as a list of publications (PubMed IDs) from which the evidence had been generated. Information for each publication was provided in the form of an XML file per publication obtained from PubMed[9]. Two sets of such data were provided initially for training (677 evidences) and development time testing (178 evidences), and an additional set was used for testing the final system (183 evidences). Bottom-line summaries with no associated abstracts and abstracts containing no text were not included in this data set. Figure 4.7 illustrates how the data in the shared task was provided. In the figure, the first column is the instance ID, the second column is the grade, and the following columns represent the PubMed IDs of the abstracts associated with each instance. All these abstracts have the same structure as the one shown in Appendix A.

```
75474 B 18492531
75475 B 12597676
75476 C 9394980
75477 B 16536797 16308411
75478 B 8582464
75479 C 10960901
15561 B 3822681 10940092 10937475 12390647 10971664 11669346
15562 B 10937475 2191938 2861907 3282670
74132 C 9215014
74133 A 11882771 9215014 9606614
74134 B 9606614
74135 B 17065896 14970960 12373695 11166969 16213659
17512 A 3147087 2040866
17513 A 2004476 7641412 8481068
```

Figure 4.7: Sample data from the 2011 ALTA shared task.

Based on the findings of our preliminary experiments, we use publication types and article titles as feature sets. In addition, we introduce word n-grams from the article abstracts as a feature set[10]. We now provide a description of these feature sets.

[9] http://www.ncbi.nlm.nih.gov/pubmed/. Accessed on 26th May, 2014.

[10] We have experimented with other features but this combination produced the best results.

N-grams

The most important information contained in the articles lies in the text of the abstracts. These include types of studies, sizes of studies, background information, results, and outcomes. To attempt to capture these information, we generate n-grams (n = 1, 2, 3 and 4) for each of the abstracts in the training set. Prior to generating the n-grams we perform some preprocessing of the text. It is common for medical concepts to have different lexical representations. For example: *hbp*, *hypertension*, and *high blood pressure* represent the same medical concepts. The concepts can further be generalised into broad categories representing classes of these concepts. In our approach, we replace specific medical concepts in the texts with generic *'sem_type'* tags. We use MetaMap[11] to identify domain specific concepts as defined in the Unified Medical Language System (UMLS)[12]. The UMLS provides a vast vocabulary of the medical concepts and also broad semantic groups into which the concepts can be classified. For example, all disease names fall under the semantic category *Disease or Syndrome (dsyn)*. Replacing each occurrence of a disease or syndrome name with the generic tag ensures that the name does not have an influence on the classifiers used and reduces over-fitting. We use the same semantic groups as Uzuner et al. [2009]: *pathological function*, *disease or syndrome*, *mental or behavioral dysfunction*, *cell or molecular dysfunction*, *virus*, *neoplastic process*, *anatomic abnormality*, *acquired abnormality*, *congenital abnormality*, and *injury or poisoning*. We also preprocess the n-grams by stemming using the Porter stemmer [Porter, 1980], lowercasing, and removing stop words.

Publication Types

We employ the approach described in the previous subsection to automatically identify the publication types of the articles. Due to the difficulty of accurately identifying all the 11 types of publications mentioned in Figure 4.3, we further condense the number of publication types and only use the ones mentioned in the previous subsection. Abstracts that do not get captured by our automatic approach are given the tag specified in the PubMed meta-data accompanying the abstract text. For articles with multiple publication types, we only keep the tag that represents the highest quality. For example, if an article was tagged as a Randomised Controlled Trial, a Clinical Trial, and a Journal Article, we only keep the Randomised Controlled Trial tag since it has the highest quality among the three types.

This approach produces a total of 23 publication types. Articles that can neither be classified by our rule-based approach nor by the associated meta-data are given the default *Journal Article* tag. For each instance of SOR in our data set, this feature set is added as a vector of the counts of each publication type.

[11] http://metamap.nlm.nih.gov/. Accessed on 26th May, 2014.

[12] http://www.nlm.nih.gov/research/umls/. Accessed on 26th May, 2014.

Titles

We generate word uni- and bi-grams from the titles, preprocess them (in the same manner as the n-grams) and use them as features. In our preliminary research, explained earlier, we found titles to be more useful for this task than journal names and publication years.

Classification

In our initial experiments, we apply the same classifiers that we used for our preliminary research. All experiments are carried out using the software package Weka[13]. Namely, we experiment with the following classifiers: Bayes Net, Support Vector Machines (SMO), K-Nearest Neighbour (IBk), Multinomial Logistic Regression (Logistic) [Le Cessie and Van Houwelingen, 1992], and C4.5 Decision Tree (J48) [Quinlan, 1993]. We also use a Naïve Bayes classifier for comparison.

Table 4.5 presents the performances of the abovementioned classifiers for each feature set, and all the feature sets combined. Consistent with our preliminary experiments, publication types prove to be the best feature set, followed by n-grams. In all these cases, both the training set and the development test (855 instances) from the ALTA shared task are used for training, and the test set (183 instances) used for evaluation. Parameter optimisation is performed using 10-fold cross validation over the 855 instances of the training and development test sets. The best performing classifier is SMO, which performs better than the other classifiers when the feature sets are combined.

In addition to the accuracy values, we add another measure, which we call the *Average Error Distance* (AED). The intent of this measure is to estimate the extent to which the predictions made by our system differs from the actual grades. For an instance, if the actual grade is A and our system predicts B, then the *Error Distance* (ED) is 1. Similarly, if, for the same instance, the prediction by our system is C, the ED is 2. For correct predictions, the ED is 0. The formula to compute AED is as follows:

$$AED = \frac{\sum_{g \in G} ED(g_p, g)}{(2 \times (N_a + N_c)) + N_b} \tag{4.1}$$

where N_a, N_b and N_c represent the number of instances with actual grades A, B and C, g_p is the predicted grade, g is the actual grade, and the function $ED(g_p, g)$ gives the ED for that instance. The formula for AED ensures that, for a given data set, the worst performing system will obtain an AED of 1, while the best performing system will obtain an AED of 0. Thus, the lower the

[13] http://www.cs.waikato.ac.nz/ml/weka/. Accessed on 26th May, 2014.

AED for a system, the more likely it is to make accurate predictions. In other words, smaller AEDs mean that the classifier predictions are *closer* to the actual predictions. The AED metric makes it easier to compare different systems on the same data set. The performance of two systems having comparable or equal accuracies may be very different in practice. For example, a system which frequently classifies C grade evidence as B is better in practice than a system that classifies the same instances as A. In other words, the errors made by the second system are *bigger*, and the grades predicted are further from the actual grades. Although the accuracies of these two systems may be equal, their AED values will indicate their relative performances. Table 4.6 presents the AEDs for the best performing classifier for the different feature sets. Table 4.7 shows the confusion matrix for this classifier.

Automatically extracted features do not provide as good a performance as manually extracted features. Thus, despite the use of more features, we are not able to obtain results as good as those from our preliminary experiments. The highest accuracy is obtained by SMO using all three feature sets (60.1%). When performing analysis on the training set to optimise various classifier parameters, we notice that a major problem in this classification process is increasing the recall and precision for the A and C classes. Attempting to improve recall significantly decreases precision and vice versa. For the best feature combination and classifier, our system obtains an average precision and recall of 0.56 and 0.51 respectively. The B class, being the majority class, has a recall of 0.89, while the A and C classes have recalls of 0.39 and 0.24 respectively. Optimising the classifier parameters to improve the recall values of the A and C classes also causes significant drops in the recall for class B. Our post-classification analysis showed that, in some cases, an instance that is correctly classified when a single feature set is used may get classified incorrectly when all the feature sets are combined, although the overall accuracy tends to increase.

Classifier	**Abstract (%)**	**Title (%)**	**Publication Type (%)**	**All (%)**
Naïve Bayes	48.3	45.2	54.9	53.5
Bayes Net	46.7	48.6	55.7	56.1
K-Nearest Neighbour	39.3	47.1	54.6	51.4
Logistic Regression	44.2	48.1	55.9	56.6
C4.5	41.0	47.5	58.5	57.4
SMO	49.7	52.5	57.4	**60.1**

Table 4.5: Individual classifier accuracies for six classifiers using all three feature sets.

Features	Average Error Distance
Abstract Text N-grams	0.332
Title Text N-Grams	0.339
Publication Types	0.300
All	0.289

Table 4.6: Average Error Distances (AED) for the best performing classifier (SMO) for various feature set combinations.

	A	B	C
A	22	25	9
B	7	79	3
C	7	22	9

Table 4.7: Confusion matrix showing number of correctly and incorrectly classified instances when all feature sets are combined and used together. The rows show the actual classes, and the columns represent the system classifications.

Combining Classifiers

Our experiments show that combining a number of features and using one classifier can only produce accuracies of approximately 60%. We intend to explore the possibility of optimising the existing classification strategy and the feature sets to further improve performance. In our approach, we apply a sequence of classifiers, each of which is provided with a specific feature set, instead of all the feature sets at the same time. The intent of each classifier in the sequence is to attempt to identify instances belonging to the A and C classes with relatively high precision, at the expense of recall. Thus, each classifier in the sequence only classifies a small number of instances as A and C classes. When a number of such classifiers are utilised, the number of correctly classified A and C grade evidences increases at each step, with a lower number of false positives for both these classes compared to the number of false positives when all feature sets are combined in a single classifier. The sequence in which the classifiers are applied and specific details about each of them are as follows:

Step 1: Classify all evidences as grade B (majority class).

Step 2: Support Vector Machines (SVM) with n-grams (n = 1, 2, 3, 4 and semantic types replaced) as features. Parameters: $c = 2.0$ and $\gamma = 0.0$. Attribute selection: using the information gain measure to select the top 400 n-grams.

Step 3: SVMs with publication types as features. For each instance, the frequency of each

publication type is used. Parameters: $c = 1.0$ and $\gamma = 0.0$.

Step 4: SVMs with titles as features. Parameters: $c = 32.0$ and $\gamma = 0.002$.

The parameters for the SVMs are tuned using the training set for training and the development time test set for evaluation. Each of the above classifiers and their parameters are chosen based on their precision in classifying A and C grade evidences. Thus, in our algorithm, each classifier classifies most instances as B but identifies some A and C class instances with high precision. At each step, instances classified as A or C are removed from the set, and the new grades are assigned to these instances. Using this approach, the classification accuracy increases with each step of the algorithm as more instances are classified as A and C.

For the final evaluation, we train our classifiers using the training set and the development test set, and evaluate the performance using test set instances. Among the 183 instances of the test set, our classifiers classified 36 as grade A, 130 as grade B, and 17 as grade C. This achieves an overall accuracy of 62.84%, meaning that 115 instances out of the 183 were correctly classified. This is significantly better than the baseline of classifying all instances as grade B, which has an accuracy of 48.63% (CI: 41.50 – 55.83). The AED value for this is 0.271, which is smaller than those of the previous experiments. Table 4.8 presents the confusion matrix for this classification strategy, and Table 4.9 presents the F-score for each class at each state of the sequence. From the table, it can be seen that the F-scores for all three classes tend to increase with each step of the sequence. The average F-score for the three classes is 0.555, compared to the average F-score of 0.506 when all feature sets are used in a single classifier. In addition to increasing accuracy and decreasing AED, an advantage of this approach is that more high precision classifiers can be plugged into this pipeline. This can further increase accuracy while also ensuring that very few A grade evidences are classified as C and vice versa.

	A	**B**	**C**
A	25	29	2
B	6	79	4
C	5	22	11

Table 4.8: Confusion matrix showing number of correctly and incorrectly classified instances. The rows show the actual classes and the columns represent the system classifications.

Class	Step 1	Step 2	Step 3	Step 4
A	NA	0.278	0.500	0.543
B	0.486	0.666	0.706	0.721
C	NA	0.094	0.415	0.400

Table 4.9: The F-scores for the three classes at each step of the sequential classifier.

4.5 Human Evaluation

In this chapter, we have so far explained our supervised classification model for the automatic grading of evidence. We showed that by sequentially applying high precision classifiers, it is possible to achieve accuracies of over 60% and also minimise the AED. We have compared our system to the majority class baseline and showed that our approach achieves significantly higher accuracies. However, to fully understand the applicability of our system in real life grading of evidence, it is pertinent to compare its performance against the performance of human experts on the same data. This is particularly important because of the following two reasons:

1. The articles in the *Clinical Inquiries* section of the JFP are authored by different domain experts. These experts follow the guidelines of evidence-based medicine to answer the clinical queries that are the topics of the JFP articles. For each article, the authors retrieve the relevant literature available on the topic, read and analyse them to identify the key recommendations, synthesise information from multiple relevant documents, appraise the quality of the information they have gathered, and then choose a grade based on the guidelines of the SORT. To specify the final grade of evidence, they combine their domain knowledge and the information available to them, and make the judgements based on them. The guidelines, as shown and explained earlier, only provide high-level instructions on the grading procedure. The authors are required to compare the information available to them with the guidelines available and make the judgements. In many cases, as expected, the information available to the authors cannot be directly mapped on to the rules of the guidelines. This may happen, for example, when there are multiple articles available on the same topic, with contrasting information in some of them. The authors may choose to refer to the *best* papers only when making the final recommendations, and ignore the lower quality articles presenting contrasting information. The final grade assigned, therefore, will depend only on the papers on which the authors chose to rely. However, there is a strong possibility that if other experts were given the same set of documents to make a decision, they would apply their own methodology to derive the final grade for the evidence. As such, there is also a strong possibility that different experts, when faced with the same clinical query and the same set of supporting information, will choose different grades to specify the qualities of the evidences, particularly when there are time-related constraints.

It is vital to compare the agreement levels among different experts in order to understand how good or bad the performance of our system is.

2. The human experts have access to more information than our system. The experts can utilise information from the full articles, while our system has to rely on information present in the abstracts only. Furthermore, in a number of cases (approximately 10%, as explained in Chapter 3), the abstracts do not contain any text, making it harder for the system to achieve its goal. Therefore, it is likely that the system's predictions are affected by the lack of information that is provided to it.

Due to these two reasons, it is difficult to estimate the actual performance of our system and the supervised classification model that we propose. To better understand the true performance of our system, it is essential to compare the grades generated by it to grades determined by human experts using the same data. Therefore, we designed an evaluation experiment involving humans.

4.5.1 Experiment Design

The intent of this experiment was to compare grades assigned by human experts, and grades generated by our system, to the grades that are used as the gold standard in our classification task. In particular, we wanted to investigate the following two issues:

i The extent to which human experts agree with the gold standard annotations in our corpus, given the same data that our system receives.

ii The extent to which human experts agree amongst themselves regarding the evidence grades, given the same data that our system receives.

To commence this experiment, we first selected a random set of 100 instances from the text classification task described in the previous section. Each instance consisted of a clinical query, a set of article abstracts that were referenced to generate the bottom-line summary associated with the evidence, and a grade indicating the quality of evidence. Among the 100 instances, 38 were of grade A, 36 were of grade B, and 26 were of grade C[14].

We employed four human experts to perform the grading task on this data set, using their own expertise. The human experts chosen were from different backgrounds, but prior to introducing them to the task, they were tested on their knowledge of evidence-based medicine practice. Three of the human experts were practising chiropractors who graduated from Macquarie University,

[14]The proportions for each of these three grade categories are not the same as in the full data set purely as a result of the random selection process.

and the fourth expert was a final year medical science student from the Australian National University. For the task, we implemented a web-based tool that enabled the experts to read the queries and the associated abstracts, and choose a grade for each evidence based on the given information. Each expert received one hour of training prior to the grading task. During the training process, the experts were introduced to the SORT guidelines, and they performed grading on a small set of examples that was separate from the 100 instances mentioned earlier. Following the grading of this sample set, the JFP grades were revealed to the experts, and the reasoning behind the grades were also explained. Using the web-based tool, the experts were able to access the instances at any time; the tool also allowed them to revise the grades they assigned. Using the tool, the experts could leave comments justifying their decisions. Figure 4.8 gives an example of the interface of the tool used by the experts to assess the grade of evidence.

4.5.2 Experiment Results

In this subsection, we present the results of our human evaluation experiments, and we discuss the conclusions we can derive from the results obtained. We compute a number of statistics, based on which we can better understand the performance of our system and the validity of our supervised classification model. Figure 4.9 shows the grade distributions for the gold standard, our system, and the four experts. The only significant distinctions between the distributions that can be noticed from the figure is the high number of B grades assigned by our system and the low number of C grades. This is due to the design of our system, as explained in the previous section.

By considering the grades in the gold standard as the correct grades, we compute the *accuracies* for the four experts and our system. Table 4.10 shows the *accuracies* obtained by the four expert graders and our system on this small data set of 100 instances. It can be seen that three of the four experts obtain better *accuracies* than our system, and one expert obtains a lower *accuracy*; only one of the expert's *accuracy* is statistically significantly better than that of our system. The results are encouraging for our system because they suggest that the performance of our approach is at least close to the performance of human experts. Also, given the same data, our system's grades and the experts' grades have similar differences with the gold standard grades. However, the results do not explain whether the reason behind this is the lack of available information, or if there is a good amount of disagreement among the different human authors on the same information. Therefore, we investigate the agreements among the different experts.

We use the Cohen's Kappa [Carletta, 1996] measure (shown below) to compute inter-expert agreements.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4.2}$$

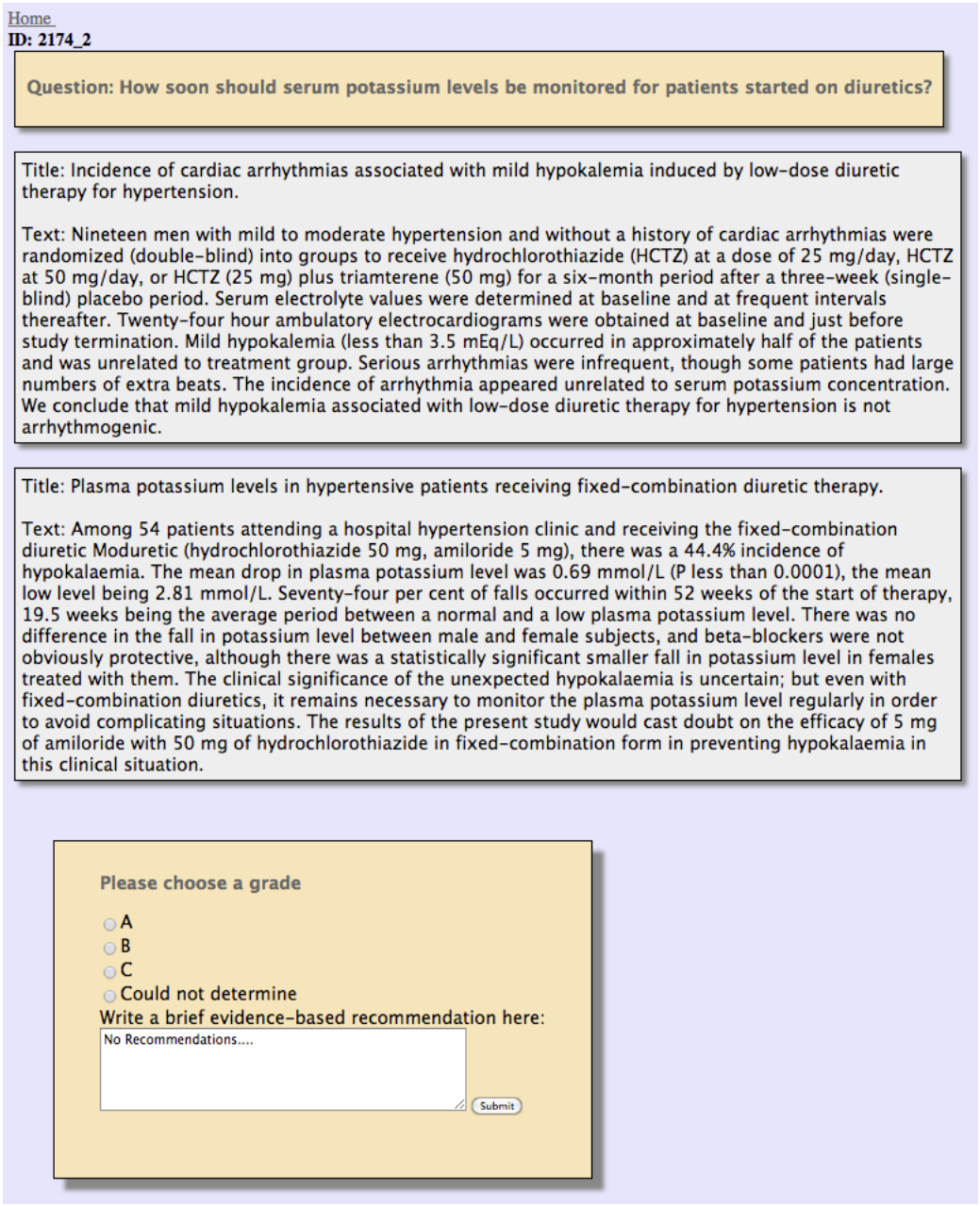


Figure 4.8: The interface of the tool used by human experts for evidence grading.

	System	Expert 1	Expert 2	Expert 3	Expert 4
Accuracy	0.61	0.58	0.69	0.71*	0.62
95% CI	0.51–0.70	0.48–0.68	0.59–0.78	0.61–0.80	0.52–0.72

Table 4.10: *Accuracy* values for the four experts and our system, along with 95% confidence intervals, when compared to the gold standard annotations. * indicates statistical significance.

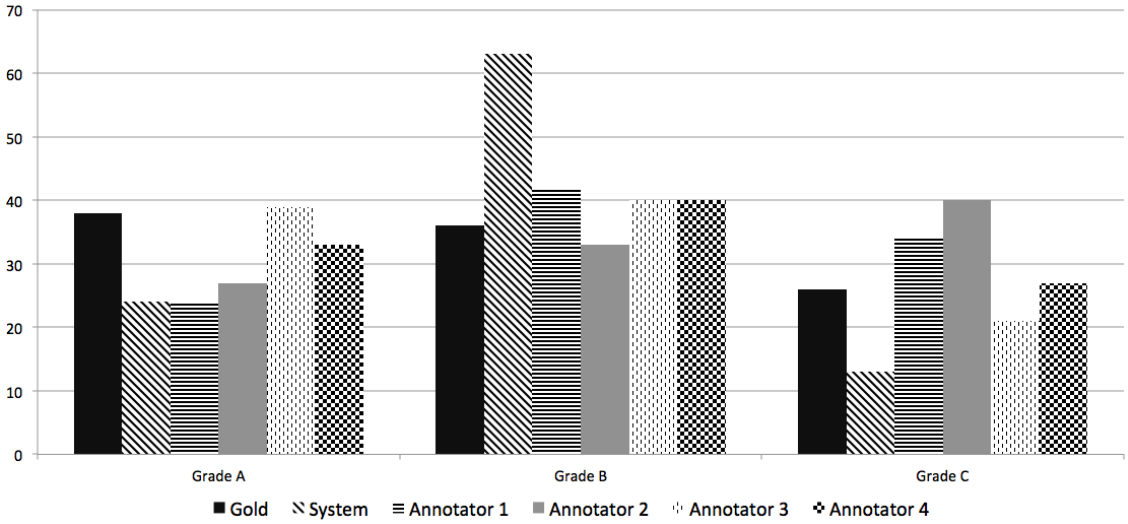


Figure 4.9: Grade distributions for the gold standard, our system, and four experts.

where $Pr(a)$ is the relative observed agreement among two experts, and $Pr(e)$ is the hypothetical probability of chance agreement. We apply this statistic in two different ways to obtain an understanding of agreement between experts. In the first approach, we compute the pairwise agreements between the expert grades and the gold standard grades. This gives us a better understanding of the extent to which the experts employed by us agree with the JFP contributors. In the second approach, we compute the pairwise agreements between the experts only, and this gives us an idea of the extent to which the experts agree amongst themselves, based on the data available to them.

Table 4.11 presents the pairwise agreements between each expert and the gold standard grades along with the mean, standard deviation and 95% confidence intervals. For all the agreement measurements presented henceforth, the confidence intervals have been computed using the formula $mean \pm 1.96 \times stdev$. From the table, it can be seen that the agreement values for Expert 1 and Expert 4 are much lower than the agreement values for the two other experts. The mean agreement is 0.47, which can be regarded as moderate agreement [Landis and Koch, 1977]. The standard deviation in agreement, at 0.09, is also relatively high. These agreement values indicate that experts only have moderate agreement amongst themselves regarding the grades. Using the same measure, our system's grades have an agreement of 0.40 with the gold standard grades. These agreement values support the idea that there is not complete agreement among the experts when it comes to assigning evidence grades. This also suggests that our system's performance is promising, given the information that it is provided.

Table 4.12 presents the pairwise agreements between the experts, along with the mean and the standard deviation. The mean agreement, at 0.52, is slightly higher than the mean agreement

	Agreement (κ)	95% CI
Expert 1	0.37	0.19 – 0.55
Expert 2	0.54	0.36 – 0.72
Expert 3	0.56	0. 38 – 0.74
Expert 4	0.42	0.24 – 0.60
Mean Agreement	0.47	
Standard Deviation	0.09	

Table 4.11: Pairwise agreements between each expert and the gold standard grades, along with the mean, standard deviations and 95% confidence intervals.

	Agreement (κ)	95% CI
Expert 1, Expert 2	0.50	0.34 – 0.66
Expert 1, Expert 3	0.42	0.26 – 0.58
Expert 1, Expert 4	0.46	0.30 – 0.62
Expert 2, Expert 3	0.54	0.38 – 0.70
Expert 2, Expert 4	0.61	0.45 – 0.77
Expert 3, Expert 4	0.60	0.44 – 0.76
Mean Agreement	0.52	
Standard Deviation	0.08	

Table 4.12: Pairwise agreements between the experts.

with the gold standard, although this increase is not statistically significant. From the table, it can be seen that Expert 1 has particularly low agreement with the other experts, and if the grades assigned by Expert 1 are ignored, the mean agreement rises to 0.58. However, the level of agreement can still be considered to be moderate [Landis and Koch, 1977]. This further supports the possibility that despite the presence of a clear and simple guideline for the SORT, experts still vary in their assignment of grades.

To further compare our system against the expert grades, we compare the AEDs of our system on this data with the AEDs of the experts' grades. Table 4.13 presents the AEDs, which are computed relative to the gold standard grades. It can be observed from the table that the AED values resemble the accuracy values, with the system's AED falling between the AEDs of the four experts' grades, and Expert 3's grades having the best (lowest) AED value. The comparable AED value of our system to the expert annotations further shows the promise of our classification model and approach.

	Average Error Distance (AED)
Expert 1	0.28
Expert 2	0.24
Expert 3	0.21
Expert 4	0.26
System	0.26

Table 4.13: Average Error Distances (AED) for the experts' grades and our system's grades.

As the final comparison between our system's grades and the experts' grades, we compute the pairwise agreements between our system and the grades assigned by the experts. Table 4.14 shows the agreement values, along with the mean and standard deviation. From the table it can be observed that mean agreement is lower than the mean agreement between the experts, and this can be considered to be *fair* agreement [Landis and Koch, 1977].

Discussion

We have obtained some interesting results from the human evaluations described in this section. We can summarise the key findings of this evaluation as follows:

1. From the agreement measurements we see that the mean κ for the agreements among the experts employed by us is marginally higher than the mean κ for the expert–gold standard agreements;

2. The *accuracies* and AEDs of our system and the experts are comparable in general; and

	Agreement (κ)	95% CI
Expert 1, System	0.35	0.29 – 0.41
Expert 2, System	0.29	0.23 – 0.35
Expert 3, System	0.35	0.29 – 0.41
Expert 4, System	0.35	0.29 – 0.41
Mean Agreement	0.34	
Standard Deviation	0.03	

Table 4.14: Pairwise agreements between the experts and our system.

3. There is significantly lower agreement between the experts and our system compared to the agreement between the experts only.

The most likely reason behind finding (1) is the fact that the JFP contributors have access to the full articles, whereas the experts employed by us only had access to the abstracts. When generating evidence grades, it may be beneficial to incorporate information from full documents, rather than just the abstracts. However, the performance of our supervised classification model and system using features derived from full articles can not be determined at this point due to the unavailability of the full articles. Furthermore, the JFP contributors are likely to review more articles than those cited. Thus, if we provide our system with all the articles retrieved, its performance is likely to change.

Finding (2) shows the promise of our supervised classification model and the approach we use for the classification task. If the experts employed by us had very high agreement or *accuracies* compared to the gold standard, that would mean our system's performance is poor. However, since our system achieves accuracies and AED values comparable to the human experts, we can conclude that it is not possible to significantly improve the accuracy of our system with the data that is currently available. We defined the score AED based on the argument that a good system should try to minimise the difference between its grades and the grades assigned by a human expert. Our system's AED score is comparable to the AEDs of the human experts, relative to the gold standard. This further suggests that our approach successfully reduces the number of cases when the generated grades are very different from the actual grades, and the *amount* of error made by our system is similar to the amount of error human experts would make, given the same data.

Finally, finding (3) shows that despite the good relative performance of our system, there is still room for modifications that can be made to our system, and which can perhaps provide marginal improvements. In other words, there is still a significant difference in the way our system predicts a grade and how experts derive their decisions. There is only little agreement among the experts

and the system, showing that future research should focus on updating the automatic grading process so that it can resemble the human grading process more accurately.

Our experiments also enabled us to estimate the time required for manual appraisal of evidence. Our human experts, on average, required approximately 11 hours to complete the grading of the 100 instances of evidence. The automatic classification approach using a trained classifier requires less than a minute to classify all the training instances in the evaluation set. Considering the performance of our system relative to human experts, and the time required for automatic classification, we can conclude that such an automatic approach has the potential of being very useful for aiding practitioners in everyday evidence-based practice.

4.6 Chapter Summary

In this chapter, we addressed the problem of automatic grading of evidence on a chosen discrete scale. We first discussed the grading scale (SORT), the various grading criteria, and the grading approach in detail, and we proposed a supervised machine learning model to solve the problem (Section 4.2). In Section 4.3, we described our analysis, which was carried out with the intent of identifying useful factors that influence evidence grades, and the suitability of a supervised machine learning model for this task. Our experiments produced significantly better results than the baseline, and suggested that supervised machine learning has the potential of being applied to this task. We also made some key discoveries regarding the importance of various factors in determining evidence grades. Specifically, we discovered that the publication types of individual articles are useful predictors of evidence grades, and the titles of articles are also useful. However, publication dates (years) and publication venues (i.e., journal names) of individual articles are not useful predictors of evidence grades according to our analysis.

Due to the importance of the information regarding the publication types of individual articles in the grade classification process, we attempted to devise an automatic approach for identifying the publication types of medical articles. We showed that a rule-based approach can efficiently identify the publication types of high quality articles such as Systematic Reviews and Randomised Controlled Trials by utilising information from the article titles, abstracts, and the associated meta-data. Automatic identification of lower quality publication types such as Case Studies is more challenging since the article titles and abstracts often do not contain the necessary information.

We applied supervised machine learning with automatically extracted features to perform the grading task. In our model, we applied a sequence of classifiers that attempted to separate A and C grade evidences from B grade ones. We obtained an accuracy of 62.84% using this approach, which was a significant improvement over the baseline. We introduced an evaluation metric called

Average Error Distance (AED), which attempts to estimate the *closeness* of a system's grades to actual grades, and we showed that our sequential classification model achieves improved AEDs compared to the baseline.

To conclude our research on this topic, we conducted a human evaluation and compared the performance of our system with human experts. Our experiments revealed that when human experts are given the same data as our machine learning algorithm, they only have *moderate* agreement regarding the grades. The experiments also revealed that although the performance of the experts is comparable to our system when compared against the gold standard, there are still significant disagreements between the expert assigned grades, and the grades assigned by our system. Based on our findings, we can conclude that supervised classification is a promising approach for automatic grade recommendations. Considering the relatively low level of agreement between human-generated and automatically-generated grades, there is still room for modifications/adjustments to the system to increase its agreement with human experts. Importantly, our evaluations suggest that it may not be possible for an evidence grading system to significantly improve on the performance using the data that is currently available.

Future research can benefit from the use of more annotated data. The amount of data used in this supervised classification model is relatively small, and this affects the performance of the classifier particularly for the smaller classes such as C. Furthermore, in our grading model, we have only incorporated features from individual documents, and we have not utilised multi-document features such as consistency. Automatic extraction of such multi-document features is challenging, and current research work in this area is limited. We have explored document level polarity classification for this domain [Sarker et al., 2011], but the accuracies of such techniques are still not sufficiently high to be applied as an intermediate step in the evidence grading task. Future improvements to such techniques may deem them suitable for use in the automatic grading task. Based on the findings of our human evaluations, however, it appears that the amount of improvement that can be achieved is limited.

Another important future research task will be to model the problem of automatic grade generation as a regression task rather than classification task. Since the three grades represent ranks for the document sets, automatic approaches may benefit from a regression based model — where the rank for a document set is predicted by a score derived via regression, rather than a specific grade. A regression model, however, will require more data since separate sets have to be used for training, rank-threshold computation (*i.e.*, finding the boundaries between the scores reflecting different grades), and evaluation.

5 Single-document, Query-focused Text Summarisation

5.1 Introduction

As previously mentioned, our model for evidence-based text summarisation involves two tasks: generating bottom-line answers to clinical queries based on available evidence, and grading the quality of the extracted evidence. In the previous chapter, we discussed the problem of generating quality grades for medical evidence, which is an important component for a summarisation system in this domain. We modelled the problem of predicting evidence grades as a supervised classification problem, and we attempted to solve it with a sequence of high precision Support Vector Machine (SVM) classifiers. We showed that, using such a classification strategy, approximately 63% accuracy can be achieved. This is significantly better than our baseline for this challenging problem, and this accuracy is also comparable to *human performance* on the same data.

In this chapter and the next one, we focus on the text summarisation component of an evidence-based, query-focused, text summarisation system. For the summarisation component of the system, we proposed the use of a two-step summarisation model. In the first step, content is selected from individual documents in response to clinical queries; in the second step, the selected content from individual documents are utilised to generate evidence-based recommendations. We focus on the first step of the model in this chapter.

We model the task of extracting information from single documents, based on the information needs of a query, as a query-focused, single-document, extractive text summarisation task. In terms of inputs and outputs, we formulate the task as follows:

Input-1: A clinical query
Input-2: A source article abstract

Output: An extractive summary of *Input-2* based on the information needs of *Input-1*

In Chapter 3, we presented samples of these target single-document summaries. Our corpus contains a collection of human-authored detailed justifications (single-document summaries), which are relatively detailed and present various types of information from different locations of the document. The content of a summary usually includes some background information — such as the type of the study, the number of subjects involved in the study — and, importantly, the outcome of the study. Figure 5.1 provides an example of a human-authored single-document summary from our corpus. In the two-sentence summary provided in the figure, the first sentence provides useful background/study details, while the second summarises the outcome. In medical document abstracts, background/study/population information tends to occur early on in the document, while outcome information is generally presented towards the end. This indicates that any summarisation algorithm attempting to automatically extract query-focused content from single documents must be capable of providing appropriate coverage of content from various document locations, instead of just focusing on extracting the outcome.

Question.
What is the best treatment for hypertension in African Americans?

Summarised Answer:
A randomised controlled trial compared the effects of consuming the DASH diet (consisting of 4-5 servings of fruit, 4-5 servings of vegetables, 2-3 servings of low-fat dairy per day, and less than 25% fat) with a typical high-fat control diet among 459 adults with normal or elevated blood pressure. Among 133 patients with hypertension, the DASH diet reduced systolic and diastolic blood pressure by 11.4 mm Hg (97.5% confidence interval [CI], -15.9 to -6.9) and 5.5 mm Hg (97.5% CI, -8.2 to -2.7) respectively when compared with the control diet. [*PubMed ID: 9099655*]

Figure 5.1: A clinical question and an expert generated summary of a medical document based on the information needs of the query.

In Chapter 2, we presented a detailed survey of extractive text summarisation techniques. We presented the *Edmundsonian Paradigm* [Edmundson, 1969] for extractive summarisation, where each text segment is assigned multiple scores based on various factors, and the final score of the segment is the weighted sum of the individual scores. Research on extractive summarisation has largely followed this paradigm. The key to successfully utilising this paradigm is the identification of important factors and the weights associated with them. Despite its simplicity, the paradigm is extremely flexible, and therefore, it has been applied in automatic extractive text summarisation

approaches in all domains. Early research works utilised word frequencies, sentence positions, keywords, and similar other surface features (superficial features within a text segment with little or no linguistic information) [Luhn, 1958, Baxendale, 1958, Earl, 1970, Kupiec et al., 1995, Hovy and Lin, 1998]. The general summarisation algorithm for a system that attempts to select *n* sentences from the source document by using *m* features is as follows:

1. Assign *m* scores to each sentence based on each of the features;
2. Update each of the *m* scores by multiplying them with predefined weights;
3. Compute the total score for each sentence by combining (*e.g.*, summing) each of the *m* weighted scores;
4. Select the highest scoring *n* sentences; and
5. Present the *n* sentences as the final summary, maintaining their order in the original document.

Although surface features are simplistic in nature, in practice they are very effective in specific domains. For example, the relative position of a sentence has been the single most important feature in the extractive summarisation of news documents till date. News articles invariably have a well-defined structure, and the most informative sentence is inevitably the first sentence. As such, the baseline of *first n sentences*, used for news summarisation tasks, has proven to be almost impossible to beat [Ceylan et al., 2010]. Therefore, extractive summarisation systems in the news domain tend to reward sentences occurring earlier in the document, when assigning scores. Likewise, in the medical domain, especially in summarisation tasks involving medical research papers as source documents, sentences occurring later in a document are assigned higher weights since later sentences are more likely to contain outcomes or conclusions of studies. Consequently, a common baseline in this domain is the *last n sentences* [Lin and Demner-Fushman, 2007].

Such a scoring approach intentionally adds bias to the sentence selection process. Our analysis of the corpus, particularly the human-authored, single-document summaries, suggests that biasing the selection of all the target summary sentences can deteriorate the summariser's performance by reducing diversity and increasing redundancy. As already illustrated, the summaries tend to present some key background information along with outcome information. While outcome information occurs towards the end of a document, background information tends to occur very early in the document. Taking the position-based scoring as an example, the algorithm mentioned above presents some obvious problems. Giving higher scores to sentences occurring later in the document reduces the probability of sentences occurring earlier in the document of being selected for the final summary, despite their possible importance. Such models, therefore, suffer from the

problem of underfitting, and this is not desirable. One way of addressing this issue is to assign a lower weight to the position-based score. However, this also reduces the chances of selecting important outcome sentences.

In our summarisation approach, we use a target summary length of three sentences. This is because our manual analysis of the human-generated summaries suggests that three sentences is the appropriate length — capable of providing sufficient content, maintaining compactness. Research closely related to ours suggests likewise [Lin and Demner-Fushman, 2007]. Adhering to the Edmundsonian Paradigm, we explore a number of features for sentence level scoring. We combine target-sentence-specific features and features independent of the target sentence number. We also use features that incorporate query focus. Furthermore, in our summarisation algorithm, we take into account the question *type*, and how the contents of answers vary based on the question type. We, thus, customise our scoring approach based on the *type* of query posed. The scores are recomputed for each target sentence selection. For each target sentence, the weighted scores for each of the source sentences are summed to compute the final score for a sentence. The highest scoring sentence for each target sentence is then chosen for the summary.

We evaluate our approach automatically using ROUGE [Lin, 2004]. We compare the ROUGE-L[1] F-scores of our system with several baselines and show that our summarisation approach outperforms the baselines with statistical significance. We also apply a percentile-rank based evaluation for our system and all the baselines, to compare them on a relative scale. The following list provides a summary of the contributions of this chapter.

i We use a corpus specialised for summarisation in this domain, and show that summarisation performance can significantly benefit from the use of statistics derived from such specialised data.

ii We apply *target-sentence-specific* scoring to diversify our sentence scoring technique during the summarisation process, and show that this approach results in better content selection.

iii We incorporate domain knowledge in various ways into our summarisation technique. We incorporate information about the *types* of clinical questions, query-sentence semantic similarities, medical semantic types and semantic associations. We show that, for a domain-specific summarisation task such as this, incorporation of domain knowledge in various ways is effective for boosting summarisation performance.

iv We apply an automatic evaluation approach that can be used to compare the relative performance of extractive summarisation systems on a common data set.

[1]ROUGE-L attempts to find the *longest common subsequence* (LCS) between two summaries, with the rationale that summaries with longer LCSs are more similar.

The remainder of this chapter is structured as follows. We first describe some notational preliminaries in Section 5.2. In Section 5.3, we briefly overview some of the literature that is relevant to our work and forms the basis for our technique. In Section 5.4, we describe our summarisation technique in detail. We discuss a number of summarisation systems in Section 5.5, against which we compare our system. In Section 5.6 we provide our evaluation technique and show that our approach outperforms well established baselines. We conclude the chapter in Section 5.7.

5.2 Notational Preliminaries

In this chapter, we will be describing various algorithms and approaches, using as fundamentals the various types of data present in our corpus. To talk about these data and concepts succinctly, we use the following notational conventions. Some of these notations have already been mentioned in Chapter 3, when describing our summarisation model. We briefly repeat them here.

Our corpus consists of a set of records, $R = \{r_1 \ldots r_m\}$. Each record r_i contains a clinical query q_i, so that we have a full set of questions $Q = \{q_1 \ldots q_m\}$. Each record r_i has associated with it a set of one or more bottom-line answers to the query $A_i = \{a_{i1} \ldots a_{in}\}$, and a grade indicating the quality of the body of evidence associated with each bottom-line answer, so that each record has a set of grades $G_i = \{g_{i1} \ldots g_{in}\}$. For each bottom-line answer of a record r_i, a_{ij}, there exists a set of detailed justifications (single-document summaries) $L_{ij} = \{l_{ij1} \ldots l_{ijo}\}$. Each detailed justification l_{ijk} is in turn associated with at least one source document d_{ijk}. Thus, our corpus has a set of source documents, which we denote as $D_{ij} = \{d_{ij1} \ldots d_{ijo}\}$.

In the research work described in this chapter, we perform query-focused, single-document summarisation. Therefore, we only use the set of questions Q, the set of detailed justifications L, and the set of referenced documents D from our corpus. We now provide further notational formalisms specific to the research work relevant for this chapter only. Each summarisation task and the evaluation of its performance involves a query q_i, the associated human-generated detailed justification l_{ijk}, and the source document d_{ijk}. We thus define a function *summ_extract()*, which takes as input the tuple (q_i, d_{ijk}) and attempts to produce an extract, ext_i, that most closely resembles l_{ijk}. Each document d_{ijk} consists of a set of n sentences $S_{ijk} = \{s_{ijk1} \ldots s_{ijkn}\}$ and a title t_{ijk}. The function *summ_extract()* performs target-sentence-specific summarisation. Therefore, when supplied with the required information along with document d_{ijk}, it assigns a score to each sentence s_{ijkx} of the document based on various features and also depending on the number of the target sentence. We discuss the scores, the features, and the sentence scoring process in the following sections.

5.3 Related Work

We first briefly review and summarise some of the literature discussed in Chapter 2. In particular, the purpose of this brief discussion is to set the context for the work described in this chapter. We briefly cover work in extractive text summarisation, text summarisation in the medical domain, and related evaluation techniques.

5.3.1 Automatic Extractive Summarisation

Edmundson [1969] defined the framework for much of the work on extractive summarisation in what is known as the Edmundsonian Paradigm [Mani, 2001]. He used a linear function of features to rank sentences for extraction:

$$Score_s = \alpha A_s + \beta B_s + \gamma C_s ... \tag{5.1}$$

where $Score_s$ is the score for sentence s, A_s, B_s and C_s are features of the sentence s, and α, β and γ are weights for each feature. Progress in summarisation research has been made more recently as numerous statistical approaches have been proposed [Barzilay and Elhadad, 1997, Lin and Hovy, 2000, Harabagiu and Lacatusu, 2005, Hovy and Lin, 1999, Lacatusu et al., 2003, Filatova and Hatzivassiloglou, 2004, Teufel and Moens, 1997, Chakrabarti et al., 2001, Ando et al., 2000, Hearst, 1994, Marcu, 1997, 1998, Hahn and Strube, 1997]. While some approaches applied machine learning techniques for summarisation [Kupiec et al., 1995, Lin and Hovy, 1997, Aone, 1999, Lin, 1999, Conroy and O'Leary, 2001, Svore et al., 2007, Schilder and Kondadadi, 2008], others focused on natural language analysis techniques [Miike et al., 1994, Marcu, 1998, 1999, 2000, Polanyi et al., 2004, Erkan and Radev, 2004, Thione et al., 2004, Barzilay and Elhadad, 1997, McKeown et al., 1998, Elhadad and McKeown, 2001, Sauper and Barzilay, 2009]. The key ideas behind some of these approaches have been discussed in Chapter 2.

5.3.2 Query-focused Extractive Summarisation

Query-focused extractive summarisation requires the selection of sentences from the source text that are most relevant to a query. One popular approach to this task is Maximal Marginal Relevance (MMR). The MMR measure [Carbonell and Goldstein, 1998] is also commonly applied to reduce redundancy, particularly for query-driven summarisation. In this technique, relevant sentences are rewarded and redundant ones are penalised simultaneously by considering a linear combination of two similarity measures. The technique, thus, produces a set of relatively non-redundant but relevant sentences in the final summary. We use a variant of this technique in our extractive summarisation approach. Other centrality and similarity metrics for query-focused

summarisation have been proposed in the literature [Alyguliyev, 2007, Erkan and Radev, 2004, Fisher and Roark, 2006, Radev et al., 2004].

5.3.3 Summarisation for the Medical Domain

As mentioned earlier in this thesis, text summarisation research for the medical domain its still very much in its infancy, and progress in this domain is hindered by the complex nature of the text and the lack of sufficient annotated data. Text processing systems in this domain generally use the Unified Medical Language System (UMLS)[2], which is a repository of biomedical vocabularies developed by the U.S. National Library of Medicine. It covers over 1 million biomedical concepts and terms from various vocabularies, semantic categories for the concepts, and both hierarchical and non-hierarchical relationships among the concepts [Aronson, 2001]. In our research work, we make heavy use of UMLS semantic types and concepts, as identified by the MetaMap [Aronson, 2001] toolbox.

In the detailed literature survey in Chapter 2, we covered some important summarisation systems in this domain. The majority of the summarisation systems in this domain are extractive in nature [Sager et al., 1994, Gaizauskas et al., 2001, Johnson et al., 2002, Damianos et al., 2002, Friedman, 2005, Cao et al., 2011]. Annotated corpora and statistical distributions have also been utilised for summarisation in the past [Vleck and Elhadad, 2010, Reeve et al., 2006a,b, 2007]. A significant amount of work on summarisation for this domain has been carried out under the broader research area of Question Answering (QA) [Terol et al., 2006, Weiming et al., 2007, Yu and Cao, 2008]. One work closely related to the work discussed in this chapter is that by Lin and Demner-Fushman [2007]. The summarisation component of their QA system is very similar to ours in terms of intent and source texts. It relies on the classification of information present in medical abstracts into PICO (**P**opulation, **I**ntervention, **C**omparison and **O**utcome elements [Richardson et al., 1995]. A machine learning classifier is used to classify text segments as *Outcomes*, as the authors argue that this is the most important content of a medical abstract. Therefore, text segments classified as *Outcomes* are presented as the final summary. One important difference between this system and ours is that the former requires clinical queries to be submitted in the form of PICO forms, rather than natural language. As such, selecting relevant content is likely to be easier for this summarisation system compared to ours. Furthermore, in their work, the queries are only used for retrieval, while the summarisation component does not take the queries into account. This makes the summaries generic (*i.e.*, same summary is generated in response to distinct queries).

Another system similar to ours, at least in terms of intent, is BioSquash [Shi et al., 2007], a system

[2] http://www.nlm.nih.gov/research/umls/. Accessed on 26th May, 2014.

that performs question-oriented extractive summarisation of biomedical documents through the use of statistical parsing, named-entity recognition, semantic role labeling, and graph generation. This graph-based extraction approach is evaluated on a set of 18 questions, and the system is shown to perform well. Niu et al. [2005, 2006] utilise polarity information in the sentences of medical abstracts for summarisation. The authors argue that polarised sentences generally contain outcomes or concluding statements within the abstracts, and show that summarisation can be improved with the use of this information. AskHermes [Cao et al., 2011] is another summarisation system under development; it incorporates information such as the *type* of a clinical query in the summary generation procedure. The answers generated by this system, however, are not bottom-line summaries and consist of paragraphs or collections of sentences.

5.3.4 Evaluation

The most important work related to automatic evaluation of summarisation systems that is relevant to our work is that by Lin and Hovy [2003] and Lin [2004]. The authors propose a set of metrics called Recall-Oriented Understudy for Gisting Evaluation (ROUGE) that have become very much the standards for automatic summary evaluation. The intent of the ROUGE measures is to find the similarity between automatically generated summaries and reference summaries, and it has been shown that ROUGE scores of summaries have a high correlation with human evaluations [Lin, 2004, Sparck Jones, 2007]. It has been previously applied to the evaluation of summarisation systems in the medical domain [Shi et al., 2007, Niu et al., 2006]. For these reasons, we use ROUGE for the automatic evaluation of our summarisation system. In particular, we use the ROUGE-L scores, which are based on the longest common subsequences between the generated and the gold summaries. Recently, Ceylan et al. [2010] have shown that ROUGE scores for extractive summaries within a domain follow a normal distribution with most combinations of sentences giving a ROUGE score very close to the mean. This leads to a long-tailed probability distribution for all ROUGE scores, meaning that a small increase in ROUGE score can indicate large increases in percentile ranks. They further explain that, in specific domains, such as the news domain, the baseline is very difficult to beat because they already have a high percentile[3]. We use a similar percentile-based evaluation for our system. Details are provided in the Evaluation section (Section 5.5).

5.4 Methods

In this section we detail our approach for the extractive summarisation task. We explain the process by which we incorporate various sentence level statistics from our specialised corpus,

[3]For the news domain, where the standard baseline is the first n sentences, the ROUGE score percentile is >99% and therefore extremely difficult to beat by any summarisation system.

and use them for summarisation.

5.4.1 Generation of Ideal Summaries

We commence our work by dividing the set of records (R) in our corpus into two subsets: R_{TRAIN} and R_{EVAL}. We use the data in R_{TRAIN} as our *training set*. That is, we use this set for the generation of statistics based on which the summarisation of unseen text is to be performed. We use the data associated with R_{EVAL} for the evaluation of the summarisation system built using statistics from R_{TRAIN}. R_{TRAIN} consists of 1,388 abstracts, each with an associated question and a human-authored detailed justification. R_{EVAL} contains 1,319 abstracts with the same associated information.

We first generate *ideal* extractive summaries from all the abstracts in R_{TRAIN}. Given a query (q_i), a human-authored single-document summary (l_{ijk}), and an abstract (d_{ijk}), we define the *ideal* extractive summary of d_{ijk} as the set of three sentences, $S_{BEST,ijk} \subseteq S_{ijk}$, from d_{ijk} that produces the highest ROUGE-L F-score when compared with l_{ijk}. To identify $S_{BEST,ijk}$ for d_{ijk}, we apply the function *best_rouge()*, which first generates a set of all possible three sentence combinations for d_{ijk}, $S_{combs,ijk}$. It then uses the *ROUGE-L* functionality of the ROUGE toolbox to identify $S_{BEST,ijk}$ for d_{ijk} as shown below:

$$S_{BEST,ijk} \leftarrow argmax_{S_{ijkx} \in S_{combs,ijk}} ROUGE\text{-}L(S_{ijkx}, l_{ijk}) \tag{5.2}$$

The *ROUGE-L()* function returns the ROUGE-L F-score for the sentence combination S_{ijkx} and the human-generated summary l_{ijk}. The three sentence combination with the maximum ROUGE-L F-score is added to S_{BEST}.

For each d_{ijk} in R_{TRAIN}, we have a three sentence combination that we consider to be the ideal extractive summary. We thus have a set of the three best sentences from each document in R_{TRAIN}, $S_{BEST} = \{S_{BEST,0} \ldots S_{BEST,1388}\}$[4], and we use this set to derive much of the required statistics. The target of our summarisation task is therefore:

> *To use statistics derived from S_{BEST} to attempt to select a set of sentences, S_{sel}, from an unseen document d_{ijk}, such that S_{sel} has the highest scoring ROUGE-L F-score, when compared to l_{ijk}, among all the three sentence combinations in d_{ijk}.*

[4] 1388 is the number of available abstracts in the training set.

5.4.2 Generation of Statistics

We have already explained that using the same statistics for different target summary sentences may not be the best approach for this domain. Our analyses show that the human-generated query-focused summaries tend to cover various *types* of medical information, which may come from various regions of the document. As such, biasing all target sentence selections using the same statistics is likely to underfit the summarisation model and only select similar type of data. In terms of sentence positions, for example, past research in the domain supports the rewarding of sentences appearing later in documents [Lin and Demner-Fushman, 2007]. However, in our three sentence summaries, we attempt to reduce intuition related bias, and rely on various likelihood measures derived from the corpus data. For the sentence position feature, for example, our intent is to select each of the three sentences based on position related likelihood measures. For each target sentence, the likelihood measures can be derived from R_{TRAIN}. We achieve our goal of applying a summarisation model that better fits the target data by employing target-sentence-specific scoring strategies. We select a summary containing a set of three sentences, $S_{summary,ijk} = \{s_{first}, s_{second}, s_{third}\}$, from document d_{ijk} using separate statistics, wherever appropriate, for each of s_{first}, s_{second} and s_{third}. We now provide a detailed description of the statistics used, starting with sentence position related statistics.

Sentence Position Related Statistics

We generate statistics related to sentence positions from R_{TRAIN} by computing the relative sentence positions of each of the three sentence combinations in S_{BEST}. For each of the three positions, we generate frequency distributions of their relative positions in the source texts. To compute the relative positions of sentences in a document, the first sentence is numbered 1, and, following that, the *ith* sentence is given the value $\frac{i}{len(d)}$, where $len(d)$ is a function that returns the length of document d in terms of sentences. The frequency distributions are normalised so that they sum to 1. Since there are three target sentence positions, there are three distributions, each corresponding to a target sentence number. The normalised frequency distributions for all three sentence positions over the training set are shown in Figure 5.2, with the histogram bin upper limits (*i.e.*, the maximum value for a sentence position contained in a histogram bin) shown in the x-axes.[5] Each distribution is compared with the distribution for that position when three sentences are randomly selected such that $s_1 < s_2 < s_3$. Compared to the random distributions, for the middle and last sentences, the peaks for the distributions of the sentences from S_{BEST} appear to be shifted slightly to the right, indicating a preference for sentences appearing later in the documents. This comes as no surprise as it has been shown that *Outcome* information in medical studies tends to occur towards the end, and generally it is the outcome information that forms the

[5]The bin size used is 0.2.

main body of the summary. However, for the first sentences, the relative positions in S_{BEST} tend to be shifted more to the left, showing that the first sentences in our *ideal* summaries are more likely to come from earlier in the documents. The graphs in the figure show that there is a subtle difference between the relative sentence positions of the sentences in S_{BEST} and the random sentences. For S_{BEST}, while the last two sentences tend to come from later in the documents, the first sentences tend to appear very early in the document. Our analyses suggested that the human summaries contain some background/population related information first, which generally appear very early in the abstracts, followed by more detailed results/outcome information, which tend to appear towards the end.

Given a document d_{ijk}, we want to assign three sets of scores to S_{ijk}. Each of these three sets of scores are calculated by taking into account the target sentence number, *tn*. When assigning a score to a sentence, our goal is to use a probability measure for the score, which depends on the target sentence number. Thus the score for a sentence s_{ijkx} with relative position x should be given by:

$$score_{s_{ijkx}} = P(s_{ijkx}|tn) \tag{5.3}$$

That is, the likelihood for a sentence with relative position x to be chosen as the target sentence *tn*. We estimate these probabilities using our frequency distributions, since the distributions represent the actual likelihoods of the values for the relative sentence positions for each target sentence number.

To assign scores based on relative sentence positions, we first create histograms of each of the three distributions for S_{BEST}. Each distribution is generated using 10 bins with bin sizes of 0.1[6]. We then normalise the histograms using the formula:

$$h_{tn}[k] = \frac{h_{tn}[k]}{\sum_{l=1}^{n} h[l]} \tag{5.4}$$

where $h_{tn}[k]$ represents the k*th* bin of the histogram for target sentence number *tn*. Then, for a sentence in a document with relative position x, we assign it a score which is equal to the value of the normalised frequency of the bin for x. Formally:

$$RP_{s_{ijkx}} = h_{tn}[k] \tag{5.5}$$

where $RP_{s_{ijkx}}$ is the score for s_{ijkx} based on its relative position x.

[6]We have experimented with other bin sizes. Using a larger number of bins does not improve performance over the training set, as the frequencies tend to get smaller for all bins. Using a smaller number of bins gives comparable performance.

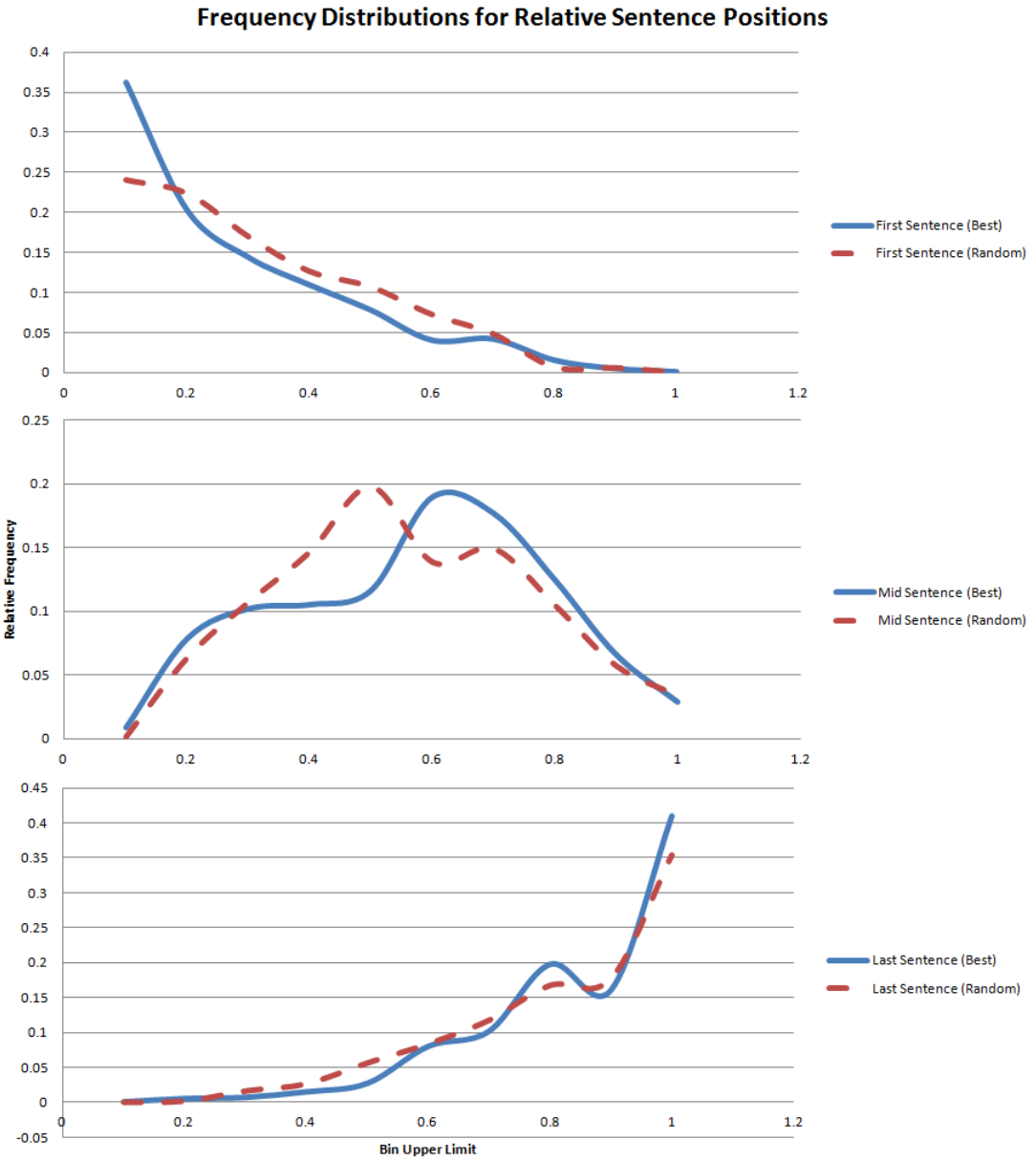


Figure 5.2: Comparison of frequency distributions of relative sentence positions for the three *best* sentences and three randomly selected sentences of each abstract.

Since we have a separate distribution for each target sentence position, the same source sentence gets a different score based on which distribution is used for scoring. Therefore, when selecting the first target summary sentence, sentences earlier in the documents get high scores. In contrast, when selecting the last target summary sentence, early sentences get very low scores. Thus, using this approach, we can assign different weights to the same sentence position depending on whether we are attempting to select the first sentence, second sentence, or last sentence of the summary.

Sentence Length Related Statistics

Our analysis of the sentences in S_{BEST} for all the sentence positions reveals that the lengths of sentences may be good indicators of their importance. On average, the sentences in S_{BEST} are longer than the rest of the sentences in R_{TRAIN}. The average sentence length for the sentences in S_{BEST} is 141.77 characters. The average sentence length for all sentences in R_{TRAIN} is 119.47 characters. Therefore, during sentence selection, we attempt to reward longer sentences and penalise shorter ones using the following equation:

$$LEN_{s_{ijkx}} = \frac{len(s_{ijkx}) - avg(len(all))}{len(d_{ijk})} \tag{5.6}$$

where $len(s_{ijkx})$ is the length, in terms of number of characters, of sentence s_{ijkx}, $avg(len(all))$ is the average sentence length in characters over the whole training set, and $len(d_{ijk})$ is the length of the document in words. The choice of using a mixture of character lengths and word lengths is intentional and purely based on the convenience of formulating the equation above. Thus, using this approach, sentences longer than the average length are rewarded by a small amount, while sentences smaller than the average length are penalised slightly. We have experimented with variations of the above equation and observed that this approach works best. This is because, for small documents, long sentences tend to contain a large proportion of important information compared to short sentences, and therefore the rewards/penalties assigned by the equation are also larger in magnitude. Using this approach, the magnitude of the score tends to be small for larger documents, and sentence selection is primarily influenced by other factors.

Sentence Similarity Related Statistics

Since our intent is to perform query-focused extraction of sentences from the abstracts, we attempt to incorporate a technique that rewards sentences similar to the associated queries. At the same time, once a sentence is selected, we try to ensure broad coverage by penalising sentences that are similar to the selected sentence. We perform this through the use of MMR and cosine similarity measures.

Similarity measures provide a single score that indicates the overlap between two sets of words: in our case, the words in the query and the words in a candidate sentence. The cosine similarity metric represents the two sets as vectors of word occurrence features. The cosine of the angle between the two vectors indicates whether a text unit is orthogonal — often loosely interpreted as “dissimilar” — to the keyword set. Treating a text unit as a vector of word features is known as a *vector-space approach* [Salton and McGill, 1983]. For most similarity metrics, including

the cosine metric, a score close to the maximum (*i.e.*, 1) indicates that there is a high overlap between reference terms and sentence words. The cosine metric is calculated using the following equation:

$$similarity = cos(\theta) = \frac{u.v}{|u||v|} \tag{5.7}$$

where u and v are two vectors, and the similarity is given by dividing the dot products of the vectors by the product of their magnitudes.

For each sentence of an abstract, s_{ijkx}, and the associated query, q_i, we generate vectors using the following approach. We first preprocess the text by lowercasing all characters, stemming the words using the Porter stemmer [Porter, 1980] and removing stop words. For each word in s_{ijkx} and q_i, we then compute the term frequency (*tf*) in that sentence and the inverse document frequency (*idf*) over all the sentences in the document. We go beyond using simple lexical similarities and incorporate domain knowledge into our approach by finding the UMLS semantic types of all the terms in each sentence using the MetaMap[7] software package. The semantic types represent broad categories of medical concepts (*e.g.*, disease or syndrome, therapeutic or preventative procedure). The intuition behind the use of semantic types, in addition to words, is that similarity in semantic types between a sentence and the query indicates that the sentence contains the same *type* of information as the query. Thus, using the UMLS metathesaurus in MetaMap, we are able to allow fuzzy matches between medical terms in s_{ijkx} and q_i when computing overlap scores. We, therefore, also compute the *tf* and *idf* measures for the semantic types in s_{ijx}. Finally, we generate vectors for each sentence using the *tf*×*idf* values for all pre-processed words and semantic types.

During extraction, for the selection of the first sentence ($tn = 1$), we compute the similarity of each candidate s_{ijkx} with the associated q_i, using the abovementioned vector representations for each. The score assigned is equal to the cosine similarity of the vectors:

$$SIM_{s_{ijkx}} = CosSim(s_{ijkx}, q_i) \tag{5.8}$$

where $CosSim()$ is the cosine similarity function defined above.

To score candidate sentences for the following two summary sentences ($tn = 2$ or $tn = 3$), we use MMR, which was introduced in page 39 and is defined by the following equation:

$$\begin{aligned} MMR_{s_{ijkx}} = {} & \lambda(CosSim(s_{ijkx}, q_i)) \\ & - (1-\lambda) max_{s_c \in S_{sel}}(CosSim(s_{ijkx}, s_c)) \end{aligned} \tag{5.9}$$

[7] http://metamap.nlm.nih.gov. Accessed on 26th May, 2014.

where $CosSim()$ is the same cosine similarity function as before, s_{ijkx} is the candidate sentence, S_{sel} is the set of sentences already selected to be in the summary, and s_c is an already selected sentence. As the equation shows, when using MMR, the highest similarity score that s_{ijkx} has with any s_c (an already selected sentence), is subtracted from the similarity score that s_{ijkx} has with q_i. The MMR score is therefore highest for sentences that are similar to the query while at the same time distinct from all other previously selected sentences. The λ parameter determines the penalty amount, with larger values favouring query similarity and smaller values favouring diversity among chosen sentences.[8]

Sentence *Type* Related Statistics

Our analysis of the human-generated summaries (l_{ijk}) in R_{TRAIN} shows that the summaries combine various types of information, such as the types of studies, the subjects involved, the design of the studies, and importantly, outcome information that are relevant to the questions. Our analysis also suggests that the sentences in a medical abstract can be roughly classified into categories that define their types. For example, a sentence may talk about the aim or intent of a study (generally at the beginning of the abstract), about the subjects involved, or the outcomes presented in the study. In our scoring mechanism, we attempt to incorporate information relating to the *type* of content present in a sentence. More specifically, we attempt to classify the type for each sentence, and assign scores to each sentence based on their types.

The first step to generate the scores associated with this metric is to categorise the sentences in our corpus. We apply the scheme proposed by Kim et al. [2011] to classify all the sentences of the abstracts in our corpus into PIBOSO (Population, Intervention, Background, Other, Study, Outcome) elements. The PIBOSO elements are a variant of the PICO elements [Richardson et al., 1995], which were initially proposed for specifying the four important components of a clinical question. Kim et al. [2011] also propose a machine learning approach for automatically classifying sentences in medical abstracts into the PIBOSO classes. In their approach, the authors use Conditional Random Fields (CRF) [Sutton and McCallum, 2007] as the learning algorithm and show that the approach works particularly well for structured abstracts. We attempt to build a more robust classification approach for the sentences.

In our approach, we divide the multi-class classification problem to six binary classification tasks, and apply one-*vs*-all classification for each. Given the features associated with a sentence, each classification task attempts to determine if the sentence belongs to one of the six PIBOSO categories and assigns a probability value to the sentence. If, for a category *C*, the probability assigned to a sentence by the classifier is greater than 0.5, *C* is considered to be a label for that

[8] We have also attempted to perform the sentence selection process in reverse (*i.e.*, selecting the last target sentence first). The difference in performance was insignificant.

sentence. If, for a sentence, the probability of $C = Other$ is greater than all the other probabilities, the sentence is labeled as *Other*.

To evaluate the performance of our sentence classification approach relative to the one proposed by Kim et al. [2011] and other baselines, we use the same corpus as the original authors of PIBOSO, and divide the corpus into two sets: training (800 abstracts) and evaluation (200 abstracts). We perform 10-fold cross validations on the training set and use it to optimise specific parameters of our SVMs. We use the 200 unseen abstracts to test the performance of our SVMs.

One advantage of formulating the problem as a set of binary classification problems is that we can customise the features to each classification task. This means that if there are features that are particularly useful for identifying a specific class only, we can use those features for the classification task involving that class, and leave them out if they are not useful for the other classes. We apply a class-specific feature set for the classification of *Outcome* sentences, which improves performance for this class.

We run MetaMap on the abstract texts to identify the medical concepts and their categories. Each medical term is assigned a *Concept Unique Identifier* (CUI) that remains unique for different lexical representations of the same concept (*e.g.*, high blood pressure and hypertension). Each medical concept is assigned a CUI and one or more semantic types by MetaMap. MetaMap also contains the MedPost/SKR parts of speech (POS) tagger, which we used to annotate each word in the data set. We further preprocess the text by lowercasing all terms, removing stop words and stemming the words using the Porter stemmer provided by the NLTK[9] Natural Language Processing toolkit. The following is a description of the features we use for the classification task:

Lexical Features. We use word n-grams as our first feature set. We have experimented with n-grams of various lengths (up to $n = 4$). Our experiments on the training set show that n-grams may be useful up to tri-grams (*e.g.*, n = 1, 2, and 3). We also add another feature set which consists of the n-grams along with the Part-Of-Speech tags for each token.

Structural Features. We use three structural features from our data set: sentence positions, sentence lengths, and section headings. The position of a sentence provides useful information about its category: a sentence towards the end of a document is much more likely to be an *Outcome* sentence than a *Background* sentence. We use both relative and absolute sentence positions as features. For structured abstracts, we use the section headings as a feature set. Section headings represent the themes of different parts of the abstracts and, therefore, provide key information about the contents of sentences belonging to the associated sections. For each sentence, we add a feature which specifies the heading of the section to which a sentence belongs

[9] nltk.org/

(if any). Our third feature in this category is the sentence length, in terms of number of words. Although previously unused for this task, our analysis show that lengths of sentences can provide useful information for specific classes.

Domain-specific Semantic Features. We incorporate domain-specific information by using the UMLS CUIs and semantic types as features. For each sentence, all the semantic types and CUIs associated with the terms in that sentence are used as features in a bag-of-words fashion (*i.e.*, ordering of terms is not taken into account).

Sequential Features. Kim et al. [2011] and Verbeke et al. [2012] both show the importance of sequential information for this classification task. Kim et al. [2011] argue that sentence level sequential information can be valuable for this task, since sentences for a particular subtopic (*e.g.*, *Background*) typically occur sequentially as a group and do not tend to repeat in later context. As such, for a sentence, incorporating information from surrounding sentences is likely to be useful for classification. Guided by the experimental results presented in Kim et al. [2011], we add, to each sentence, the preprocessed n-grams (n = 1, 2, and 3) of the previous sentences. We have experimented by including up to three previous sentences but found that only two previous sentences are useful (the second previous sentence may only be very marginally useful).

Class-specific Features. We explore the use of class-specific features for the *Outcome* class only. We incorporate a set of cue phrases from Niu [2007] that are used to identify sentences presenting outcomes. In the *Outcome* classification task, the count of those specific cue phrases present in a sentence is added as a feature.

PIBOSO Classification Accuracy

To execute the learning and classification tasks, we use the LibSVM[10] implementation for SVMs. We use an RBF kernel and optimise the c parameter using 10-fold cross validation over the training set. Table 5.1 presents the F-scores for each class for the 10-fold cross validations (CV) over the training set, and over the evaluation set. F-scores for both structured and unstructured abstracts are shown, along with the class frequencies. Our training data set contains 304 structured abstracts and 496 unstructured abstracts, while the evaluation set contains 80 structured and 120 unstructured abstracts. The table enables us to compare our results with those obtained by Kim et al. [2011] and Verbeke et al. [2012] – two benchmark systems for this task. It can be seen that the micro-averaged F-scores of our system are comparable to both the benchmark systems, and that our system tends to perform better for the less-frequent classes (*e.g.*, *study design*). The F-scores for our 10-fold cross validations for training and evaluation sets values are comparable for frequent classes (*e.g.*, *outcome*). However, for low-frequency classes, the F-scores vary more between our training and evaluation sets. This can be attributed to the small

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Accessed on 26th May, 2014.

number of instances available, and also the low agreement among human annotators for these classes, as depicted by Kim et al. [2011]. Importantly, our approach almost invariably performs better than both benchmark systems for unstructured abstracts, illustrating its low reliance on the discourse structure of documents. Notably, our system performs better despite using only 800 abstracts for the 10-fold cross validation task, compared to the benchmark systems' use of all 1,000 abstracts. We also empirically verified that our system outperforms other similar baselines. A baseline with a good performance is a multi-class SVM classifier with the same features. It achieves an overall micro-averaged F-score of 0.752, compared to our system's micro-averaged F-score of 0.766. Our sentence classification system participated in the 2012 ALTA shared task[11] and achieved second position.

Class	10-fold CV		Eval. Set		Kim et al.		Verbeke et al.		Class Freq.
	S	**U**	**S**	**U**	**S**	**U**	**S**	**U**	
Population	45.0	**59.9**	51.8	54.0	**56.3**	39.8	35.6	21.5	809
Intervention	29.8	**38.9**	**27.2**	24.7	20.3	12.9	26.1	16.1	687
Background	85.4	75.8	83.4	73.6	81.8	68.5	**86.2**	**76.9**	2,557
Outcome	91.1	**81.3**	90.8	80.3	92.3	72.9	**93.0**	77.7	4,523
Study	52.6	**60.3**	**59.6**	46.1	43.9	4.40	45.5	6.67	233
Other	87.7	**48.4**	**88.2**	40.6	70.0	24.3	88.0	24.4	3,396
Micro-average	84.1	**72.6**	**84.6**	71.6	80.9	66.9	84.3	67.1	

Table 5.1: Classification F-scores and their micro-averages for each of the 6 classes. Scores for structured (S) and unstructured (U) abstracts are shown separately for each class.

We classify all the sentences from the abstracts in our corpus using this approach, due to its good performance on the sample annotated data. Classification of all the sentences allows us to further verify our claim about the contents of the human-generated summaries. Consider the human-generated summary shown in Figure 5.3. The figure shows the best three sentence summary corresponding to the human-generated summary, with their PIBOSO classifications. It can be seen that the first sentence of the three sentence summary is classified as *Population*, and the last two sentences are classified as *Outcome* by our system.

To apply the sentence *type* information in summarisation, we analyse the distributions of these PIBOSO elements in R_{TRAIN} and S_{BEST}. Our target is to use these distributions to make probability estimates and improve the content of our summaries. We begin by generating five frequency distributions of PIBOSO elements:

- for the set of all sentences in the training set (S_{TRAIN}) (1);
- for all best sentences of the training set (S_{BEST}) (2);

[11] http://www.alta.asn.au/events/sharedtask2012/. Accessed on 26th May, 2014.

Question:
What drugs are best for bipolar depression?

Human-generated Summary:
In an RCT, 509 patients were randomized to double-blind treatment with quetiapine (300 or 600 mg/day) or placebo and assessed weekly with the MADRS and Hamilton Depression Rating Scale (HDRS). Improvements in mean HDRS scores were greater with both quetiapine doses than with placebo (P < .001). Clinical effect sizes were moderate at week 8 (0.61 at 300 mg/day and 0.54 at 600 mg/day), although dose titration effects were not established.

The three *best* sentences:

1. *This study evaluated the efficacy and tolerability of quetiapine monotherapy for depressive episodes in patients with bipolar I or II disorder (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition) who were randomized to 8 weeks of double-blind treatment with quetiapine (300 or 600 mg/d; once daily, evening dosing) or placebo. (Population)*
2. *Therapeutic effect sizes at Week 8 were 0.61 and 0.54 for quetiapine 300 and 600 mg/d, respectively. (Outcome)*
3. *This study demonstrates that quetiapine monotherapy is an effective and well-tolerated treatment for depressive episodes in bipolar disorder, confirming the results observed from a previous study. (Outcome)*

[*PubMed ID: 17110817*]

Figure 5.3: A human-generated summary and the three sentences from S_{BEST} associated with the summary. The PIBOSO classification of the sentences are shown in parentheses.

- for all 'first' sentences from the best sentences ($S_{BEST,1}$) (3);
- for all 'second' sentences from the best sentences ($S_{BEST,2}$) (4); and
- for all the 'last' sentences from the best sentences ($S_{BEST,3}$) (5).

We normalise all the frequency distributions by dividing each bin value by the sum of all the bin values for that distribution (as shown in equation 5.4). The normalised frequency distributions are shown in Figure 5.4.

The frequency distributions indicate what content from the source texts are generally included in the summary. It can be seen that the proportion of *Population*, *Intervention* and *Background*

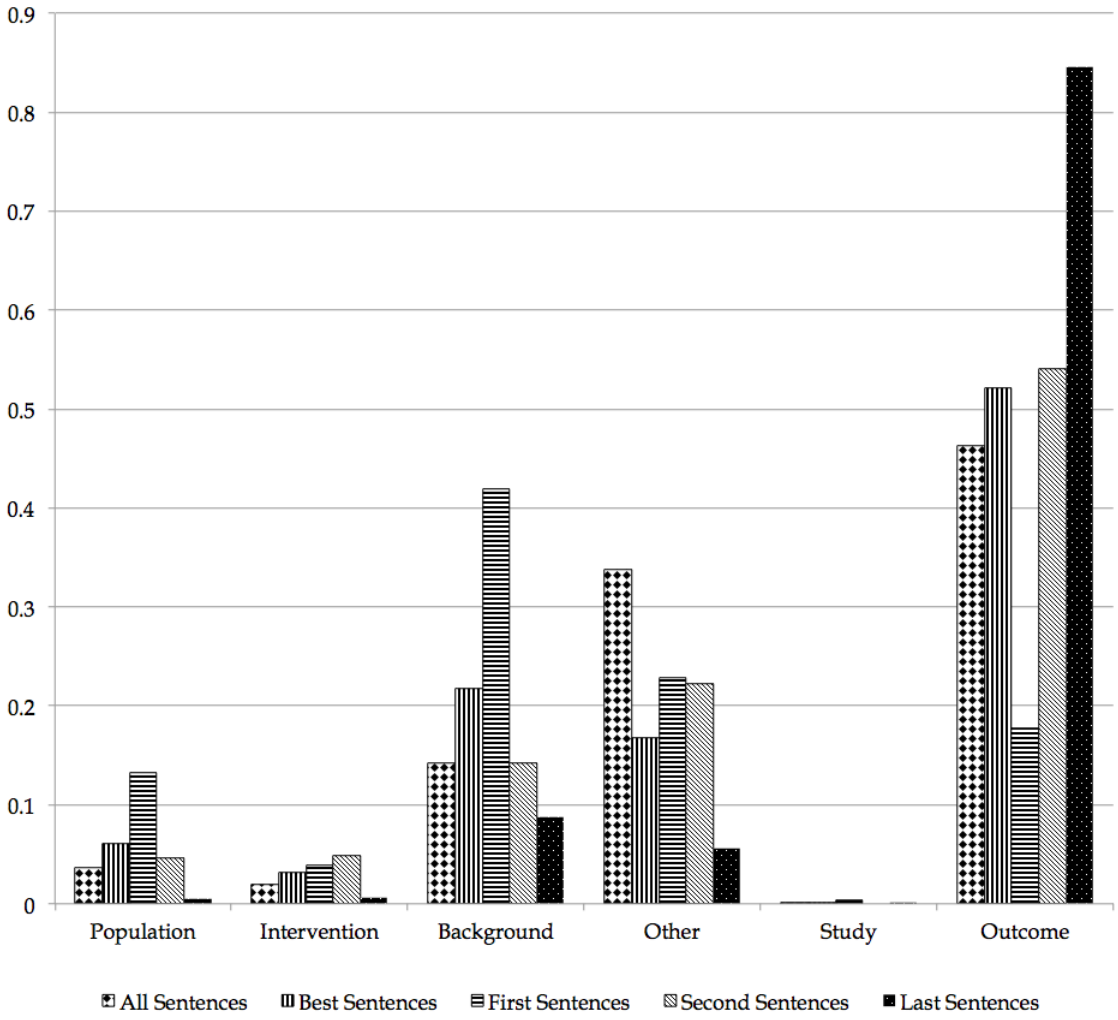


Figure 5.4: Normalised frequency distributions of PIBOSO elements over the whole training set, the best sentences, the first sentences of the best sentences, the second sentences of the best sentences, and the last sentences of the best sentences.

sentences are higher for the frequency distribution (2) compared to frequency distribution (1) and even higher for frequency distribution (3). This clearly indicates that these three types of sentences have a high probability of being chosen as the first summary sentence. Sentences classified as *Other*, in contrast, have a lower proportion for distribution (3) than distribution (1). *Outcome* sentences have a very high proportion for distribution (2), which indicates that they should be favoured when selecting sentences. However, the proportion of *Outcome* sentences for the first sentences is quite low, indicating that they are relatively unlikely to be chosen as the first sentence.

Using these frequency distributions, we can derive several likelihood estimates. For example, the likelihood of an *Outcome* sentence being in the final summary ($P(s_{Outcome}|S_{BEST})$) can be

estimated by using its normalised frequency distribution in (2). Similarly, the proportion of *Outcome* sentences in (1) provides an estimate of the likelihood for any sentence of being an *Outcome* sentence ($P(s_{Outcome}|S_{TRAIN})$). However, it is obvious that $P(s_{Outcome}|S_{BEST})$ heavily depends on $P(s_{Outcome}|S_{TRAIN})$. That is, the higher the value of the latter, the higher is the value of the former. Thus, using these probability estimates directly for scoring adds bias. So, we derive an equation that mitigates this bias, as explained later.

Target-sentence-specific probability estimates can also be made based on these distributions. For example, the likelihood of an *Outcome* sentence being chosen as the first target sentence ($P(s_{Outcome}|tn = 1)$) is estimated by the proportion of this type of sentences in distribution (3). Thus, for each type of sentence, we can compute the likelihood estimates for that type of sentence to be chosen as the first, second or third target sentence.

Taking into account the PIBOSO category of each candidate sentence, we derive two scores from these frequency distributions — one that is dependent on the target sentence number and one that is not.

The first score, which we call the Position Independent PIBOSO Score (PIPS) is computed as follows:

$$PIPS_{s_{ijkxt}} = \frac{P(s_t|S_{BEST})}{P(s_t|S_{TRAIN})} \tag{5.10}$$

where $P(s_t|S_{BEST})$ is the proportion for PIBOSO element t among the sentences in S_{BEST}, and $P(s_t|S_{TRAIN})$ is the proportion of that PIBOSO element among the sentences in S_{TRAIN}. Thus, this score is higher for sentences belonging to PIBOSO categories that have a higher proportion among the best sentences compared to all sentences; and the larger the difference between the two proportions, the higher the magnitude of this score. This score is independent of the target sentence position.

The second score is computed as follows:

$$PDPS_{s_{ijkxt}} = \frac{P(s_t|tn = x)}{P(s_t|S_{BEST})} \tag{5.11}$$

where $P(s_t|S_{BEST})$ is as before, and $P(s_t|tn = x)$ is the proportion for PIBOSO element t in the distribution (3), (4) or (5), depending on the value of tn. We call this the Position Dependent PIBOSO Score (PDPS). Thus, when selecting the first sentence (*i.e.*, $tn = 1$), a sentence classified as *Background* is given a much higher score compared to a sentence classified as *Outcome*. Similarly, when selecting the last sentence, *Outcome* sentences receive a much higher score compared to sentences belonging to other categories.

There can be a possible of six *PIPS* score as there are six types of sentences. We normalise these scores by dividing each score by the sum of all six scores. Similarly, for each target sentence, there can be a possible of six *PDPS* scores, giving a total of 18 possible scores (six for each tn). We normalise these scores in the same way.

5.4.3 Incorporating Question Information in Extractive Summarisation

Our analyses suggest that the content of a summary is influenced by the type of question. For example, the content of the answer to a question that asks about the treatment of a disease is generally different to that of a question that asks about a diagnostic procedure. Answers to the treatment type questions usually present invasive or non-invasive techniques such as drug therapy or surgical intervention. In contrast, questions focusing on diagnostic procedures may mention imaging techniques or tests, which are not associated with treatments. Our intent is to categorise the questions in our corpus into *types*, analyse the medical concepts that are prevalent in the answers to each of the question types, and devise a technique that rewards sentences by taking into account the question type and the domain specific concepts present in the sentence. In this manner, we add further query-focus to our summarisation technique. As such, we define two new scores: $ST_{s_{ijkx}}$ and $ASSOC_{s_{ijkx}}$. $ST_{s_{ijkx}}$ is a score assigned to s_{ijkx} based on the UMLS semantic types it contains and the type of q_i, and $ASSOC_{s_{ijkx}}$ is assigned based on the associations the semantic types of s_{ijkx} have with the semantic types of q_i.

Classifying Corpus Questions

We first classify the questions in our corpus into general medical topics. To do this, we apply the approach proposed by Yu and Cao [2008]. The authors use twelve separate classes for the classification task. Using n-grams, UMLS semantic types and CUIs as features, the authors attempt to solve the classification problem using supervised machine learning. A corpus, containing over 4,500 questions that were manually classified into the twelve categories by practitioners, is used by the authors for training the classifiers. The overall classification task is divided into twelve separate binary classification tasks, each of which attempts to identify a specific type of question. For each binary classification task, all questions belonging to the target category are included along with an equal number of randomly selected questions belonging to other categories. Performance is measured via 10-fold cross validation using several classifiers. In their approach, the authors obtain best classification accuracies using SVMs. We set up our classification technique and train our classifier using the same data set as Yu and Cao [2008] and use it to classify the questions in our corpus. Evaluation of the classification approach via 10-fold cross validation on their data set did not show any significant differences in performance. The distribution of question types in our corpus, as estimated through their automatic classification, is

Topic	Frequency	Proportion
Treatment and Prevention	193	0.423
Pharmacological	146	0.320
Management	135	0.296
Diagnosis	109	0.239
Test	73	0.160
Procedure	32	0.070
Prognosis	23	0.050
Physical Finding	23	0.050
Epidemiology	16	0.035
Etiology	10	0.022
History	7	0.015
Device	6	0.013

Table 5.2: Question types and their proportions in our corpus.

shown in Table 5.2. The table shows that *Treatment and Prevention* is the most frequent category, while *Device* is the least frequent. Note that each question type can have multiple categories or none. In our corpus, 216 questions have a single category, 167 have 2 categories, 61 have 3, and 9 have 4 categories. Three questions were not assigned any categories.

Using Semantic Types for Scoring

We use the categorised questions of our corpus to identify the UMLS semantic types that are important for each type of question. We then apply an additional score to each sentence s_{ijkx} based on the semantic types it contains.

Preliminary Analysis

To assess if incorporating question *type* information in our summarisation task is likely to improve performance, we performed preliminary analysis on two question types: (i) *Treatment and Prevention*, and (ii) *Diagnosis*. We first manually identified these two types of questions, analysed the semantic types that frequently occur in the answers to these types of questions, and identified a set of *important* semantic types for these two types of questions. In our analysis, during the summarisation process, the sentences belonging to these two sets of questions received an additional score. Each sentence, s_{ijkx}, received a score that is equal to the number occurrences of the identified semantic types in that sentence divided by the total number of terms in that sentence. We computed the ROUGE-L F-scores for these two sets of questions before and after adding the semantic type scores and compared the performance. For *Treatment and Prevention* questions, the ROUGE-L F-score increased from 0.1619 to 0.1644 once this new information was incorporated. Similarly, for *Diagnosis* questions, the ROUGE-L F-score increased from 0.1343

to 0.1362. The improvements indicate the positive effect this new score has on the extractive summarisation task.

Scoring Approach

Similar to some of our previous scoring approaches, we rely heavily on frequency distributions for generating a score based on the UMLS semantic types contained in a sentence. We commence by generating a frequency distribution of all the UMLS semantic types present in the human-generated summaries (l_{ijk}) of R_{TRAIN}, and normalise the frequency distribution as already shown in equation 5.4 (page 135). This gives us a measure of how the semantic types are distributed over all the different types of questions. Thus, we can say that the probability estimate of semantic type $P(st)$ of being in the final summary is the same as the normalised frequency of *st* in the generated distribution. Next, we generate separate frequency distributions of l_{ijk} for each question type. This presents us with a measure of how the semantic types are distributed for each type of question. Thus, for a question type *t*, the probability estimate of a semantic type $P(st|t)$ of being in the final summary is the same as the normalised frequency *st* in the generated distribution for *t*. Our intent is to identify semantic types that are more likely to occur for a specific type of question than for other types of questions. If a semantic type *st* has a high frequency in the distribution for question type *t*, but a low frequency in the overall distribution, it indicates that *st* is an important semantic type for answers to all questions of type *t*.

Each semantic type for each type of question has a score. The score for a semantic type st for question type *t* is calculated using the *semtype_score()* function which is as follows:

$$semtype\_score(st, t) = \frac{P(st|t)}{P(st)} \tag{5.12}$$

where $P(st)$ and $P(st|t)$ are computed as described in the previous paragraph. Thus, the *semtype_score()* is large for semantic types that are more frequent to question type *t* than the whole training set and vice versa. Once all the semantic type scores are calculated for a specific question type, the scores are normalised so that they sum to 1.

When scoring the sentences of an abstract, each sentence receives a score, $ST_{s_{ijkx}}$, based on the set of semantic types ($SEMTYPE_{s_{ijk}}$) it contains. This score is simply the sum of the normalised $semtype\_score()$ for the semantic types contained in that sentence, as shown below:

$$ST_{s_{ijkx}} = \sum_{st \in SEMTYPE_{s_{ijkx}}} semtype\_score(st, t) \tag{5.13}$$

Since the scores are normalised, the maximum score that a sentence can have for a specific question type is 1 (*i.e.*, if the sentence contains all the possible semantic types), and the minimum score is 0 (*i.e.*, if the sentence contains no semantic types). We combine this score with the

sentence level scores previously discussed.

Using Associations for Scoring

In our question-specific scoring approach, we apply another score to each sentence which we call the *association score*. The intuition behind this score is that medical terms in the questions generally have some relationship with the terms in the summary sentences. For example, if a question has a term representing a disease and the summary contains a term that acts as the cure for a disease, we can assume that there is a *is_treated_by* relationship between the disease term and the cure terms. In the medical domain, the disease and cure terms are represented by the semantic types. The UMLS semantic network also provides associations between semantic types. For example, the *dsyn* semantic type (representing *disease or syndrome*) has a 'treats' relationship with the *phsu* semantic type (representing *pharmacological substance*). We attempt to use these associations to identify sentences in the source texts that are related to the associated questions. We therefore assign a score to each sentence based on the associations its semantic types have with the semantic types of the associated question. Figure 5.5 provides an example of a semantic association between a question and a sentence semantic types. In the partial example provided, the term *fluoxetine* has a *prevents/treats* association with the term migraine[12]. We identify and utilise these associations for sentence scoring.

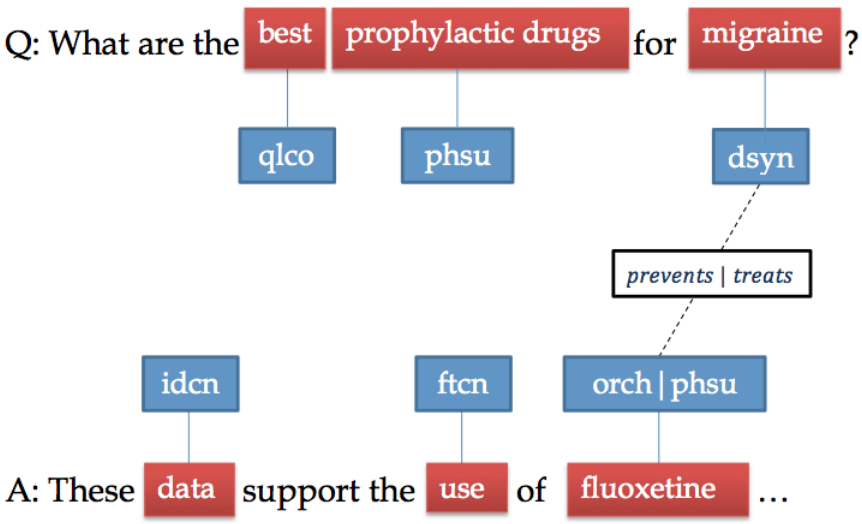


Figure 5.5: Example of association between question and summary sentence semantic types. Only the partial sentence is shown for simplicity.

[12]Note that the full example has more associations with the question. We only show one association for simplicity. In general, there are many associations between the question and sentence semantic types.

Preliminary Analysis

Our preliminary analyses of the training set data suggested that a standard sentence generally has a number of associations with the question, some of which may be useful while some may not be useful. The extent to which an association between two semantic types is useful depends largely on the information needs of the question, and hence the type of question. For example, if the question is a *Treatment and Prevention* question, which generally asks for interventions to diseases or syndromes, the association *treats* appears to be a very useful association. Such associations are very likely to occur frequently in *Treatment and Prevention* questions but not so frequently in some other types of questions such as *Diagnosis*. As such, the importance of an association varies with the type of question. In our approach, we attempt to identify the important associations for each type of question, score sentences based on the associations they have with the questions and use that score as part of the sentence ranking process.

Identifying Important Associations

To identify important associations for each type of question (*e.g.*, *Treatment and Prevention*), we commence by identifying two sets of semantic types for each type of question: (i) important question semantic types and (ii) important answer semantic types. (i) is identified from the questions in R_{TRAIN}, while (ii) is identified from the human-generated summaries (l_{ijk}) in R_{TRAIN}. We apply an approach that is similar to the one we use to identify the semantic type scores explained in the previous subsection. Ideally, for each question type, we want to identify semantic types that occur with high frequency for that type of question, while not so frequently for others. So, to identify (i), we compute semantic type frequency distributions for each type of question in R_{TRAIN}. At the same time, we compute the semantic type frequency distributions for the set of all other question types combined. We normalise both sets of distributions so that all the relative frequencies sum to 1. To ensure that our distribution does not contain any rarely occurring semantic types, we remove those semantic types that have relative frequencies below a given threshold (in our case, we empirically choose 0.01 as the threshold, which ensures that all the semantic types that are used constitute at least 1% of all the semantic types for that category). We compute the importance of each semantic type for a specific question type, by dividing its relative frequency for that question type by its relative frequency for all other question types. Finally, certain semantic types that occur frequently but are not useful for our task are removed via manual screening. For example, the *qnco* semantic type (*quantitative concept*) represents numbers which occur quite frequently but are generally not useful for our task. Appendix B presents the important semantic types for each question category.

The semantic types for (ii) are identified in an identical fashion. When identifying these semantic types, the associations (if any) between the question semantic types and the answer semantic types are not taken into account. Once both sets of semantic types are identified, we study

the important associations that exist within a question type by applying yet another frequency distribution. For each question type *t*, we compute a frequency distribution of all the associations between the important semantic types appearing in each question and the important semantic types in the corresponding answers using the UMLS semantic network.

Computing Sentence Association Scores

We normalise the association frequency distributions to obtain relative frequencies that sum to 1. Given a question q_i of type *t*, the probability estimate of the answer to that question having an association $assoc$ is the relative frequency of $assoc$ in the association frequency distribution for *t*. When scoring each sentence s_{ijkx}, we identify the set of all associations ($AS_{s_{ijkx}}$) the sentence has with the question, find the relative frequencies of the associations in the association frequency distributions and sum the relative frequencies. We use the function $assoc\_freq(assoc, t)$, which, given an association type and a question type, computes the relative frequency of *assoc* for *t*. The score assigned to the sentence is the sum of the relative frequencies, normalised by dividing the value by the total number of unique semantic types present in the question and the sentence. For questions that are assigned multiple types, the association frequency distributions for all the types are combined and normalised before computing sentence scores. The following equation summarises the scoring mechanism, which is similar to the *jaccard similarity* measure (shown in equation 5.16):

$$ASSOC_{s_{ijkx}} = \sum_{assoc \in AS_{s_{ijkx}}} \frac{assoc\_freq(assoc, t)}{|st_{q_i} \cup st_{s_{ijkx}}|} \tag{5.14}$$

where st_{q_i} and $st_{s_{ijkx}}$ represent the semantic types present in the question and the sentence being scored respectively, and $ASSOC_{s_{ijkx}}$ is the score assigned to the sentence based on the associations present. This score is combined with the previously described scores when performing the summarisation.

5.4.4 Combining Statistics for Sentence Extraction

Table 5.3 provides a summary of the features we use for the summarisation task. In the table, the *Target-sentence-specific* column indicates if the feature generates scores specific to the target sentences, the *Query-focus* column indicates if the feature incorporates information from the query, and the column *Domain-specific information* indicates if the feature incorporates domain knowledge in some way.

Feature	Target-sentence-specific?	Query-focus?	Domain-specific Information?
Relative Position	yes	no	no
Length	no	no	no
PIPS	no	no	yes
PDPS	yes	no	yes
MMR	yes	yes	yes
Semantic Type	no	yes	yes
Association	no	yes	yes

Table 5.3: Summary of the features used for the summarisation task.

We use the following *Edmundsonian* equation to give the overall score for a sentence s_{ijkxt}:

$$\begin{aligned} SCORE_{s_{ijkxt}} = \alpha RP_{s_{ijkxt}} + \beta LEN_{s_{ijkxt}} + \gamma PIPS_{s_{ijkxt}} + \delta PDPS_{s_{ijkxt}} \\ + \epsilon MMR_{s_{ijkxt}} + \zeta ST_{s_{ijkxt}} + \eta ASSOC_{s_{ijkxt}} \end{aligned} \tag{5.15}$$

where $SCORE_{s_{ijkxt}}$ is the score for a candidate sentence s_{ijkxt}, and *ijk* represents the document number, *x* represents the sentence position, and *t* represents the type of the sentence. $SCORE_{s_{ijkxt}}$ is calculated as the weighted sum of the individual scores. Note that when extracting the first sentence, we replace the MMR score with the cosine similarity score in the equation. In the case of ties, the sentence with greater length is chosen.

The weights associated with the scores let us modify their contributions towards the overall score. To automatically find good approximations for optimal values of the weights (α, β, γ, δ, ϵ, ζ and η), and the λ parameter in MMR, we perform a grid search through all values from 0.0 to 1.0 using step sizes of 0.1. Our intent is to find a combination of weights that increases the chances of selecting the sentences from an abstract that belong to S_{BEST}. Therefore, for each combination of weights obtained during the exhaustive search, we compute the recall values for the first, second and last sentences over the whole training set. The combination producing the best combined recall is chosen.[13]

5.4.5 Alternative Sentence Weighting

In addition to applying the previously mentioned approach involving an exhaustive search over the training set to learn weights for equation 5.15, we apply an alternative regression based approach for comparison. In this approach, separate weights are learned for each target sentence

[13]The grid search involves 10^8 computations and as a result it requires a large amount of computation time. However, it only requires to be run once to identify the best weights.

System	α	β	γ	δ	ϵ	λ	ζ	η
QSpec (grid)	0.8	0.5	0.3	0.2	1.0	0.6	1.0	0.3
Regression (first sent.)	0.000	0.420	0.044	0.019	0.073	0.600	0.130	0.052
Regression (second sent.)	0.001	0.016	0.035	0.074	0.019	0.600	0.300	0.084
Regression (third sent.)	0.003	0.008	0.005	0.006	0.040	0.600	0.026	0.071

Table 5.4: Feature weights for different versions of our extractive summarisation system.

using an SVM regression algorithm. In this approach, for each sentence, all the abovementioned scores are derived (taking into account the number of the target sentence) along with an additional score for the degree of overlap between the sentence and the associated human summary (l_{ijk}). Our intuition is that the higher the overlap score, the more likely is the sentence to be in the final summary. The target, therefore, is to identify weights that maximise the overall overlap scores via regression. The overlap score is calculated using the *jaccard similarity* measure given as:

$$jaccard\_similarity(S_i, S_h) = \frac{|S_i \cap S_h|}{|S_i \cup S_h|} \tag{5.16}$$

where S_i is a term vector from sentence i of source document S, and S_h is the term vector representing the human summary. Table 5.4 shows the weights used in our two scoring approaches. Note that when computing the weights via regression, each target sentence selection will apply different weights. We could do the same with our grid search technique, but choose to use a single set of weights for simplicity.

5.5 Extractive Summarisation Evaluation

We evaluate our approach automatically using the ROUGE evaluation tool. We are interested in assessing the quality of the ROUGE-L F-scores generated by our system relative to a set of baselines. Our system summaries and the baseline summaries are all evaluated by comparing with the associated human-generated summaries. The summaries we generate are not restricted by word limits. It is therefore possible to have very long sentences in the summaries. Long sentences are likely to have high ROUGE-L *recall* scores since they are more likely to cover more terms that are present in the human summaries. At the same time, they are likely to have more irrelevant content as well and, thus, have low *precision* scores. We are interested in generating summaries that have high content coverage and at the same time are not extremely long. For this reason, we use F-scores for summary quality comparisons.

Since we want to assess the relative performance of our system, we use a percentile-rank based approach for evaluating the qualities of various ROUGE scores using the technique proposed

by Ceylan et al. [2010]. To do this, we first generate the ROUGE-L F-scores for all possible three sentence combinations from the abstracts in R_{EVAL}. For each abstract, we then generate a histogram of all the three sentence ROUGE-L F-scores using 1,000 bins between 0 and 1. We normalise the histograms using equation 5.4 and use the normalised distribution as an approximation for the probability density function (*pdf*) for the ROUGE-L F-scores of the sentence combinations for the abstract.

The *pdf*s for all abstracts in the evaluation set are convolved together to generate a *pdf* for the whole set (R_{EVAL}). We, however, use a slightly modified algorithm to the one presented by Ceylan et al. [2010]. This is because we notice that, when using their approach, due to the large number of computations involved, the final convolved *pdf* can be very slightly different based on the ordering of the histograms during convolution. Our investigations suggest that the root of this problem lies in the way floating point numbers are represented in computers. Since the histogram generation and convolution operations involve millions of floating point computations, it is not strange that, over the course of the convolution process, the final generated histogram values are slightly different every time the ordering is changed. This variation is very minute, and we address this issue by executing the convolution multiple times with different orderings and then averaging the final *pdf*s to give a single *pdf*.

Figure 5.6 shows the *pdf* obtained for all abstracts in R_{EVAL}. The *pdf* shows the range of possible scores an extractive summarisation system can have given this data set. The height of the distribution at a specific score indicates the likeliness of a system of achieving that score. The distribution is long-tailed, meaning that the scores for most of the extracts in the summary space are clustered around the mean. This suggests that most systems are likely to produce scores that are around the mean of the *pdf*. According to the distribution, the minimum score that a summarisation system in this domain can have is 0.042 and the maximum is 0.255. The two ends of the distribution are shown on Figure 5.6 via the short red lines. The longer red line shows the score achieved by our system, and we will discuss this score later. However, 95% of the scores will lie within a very small range — between the values 0.139 (approximate percentile rank of 2.5%) and 0.169 (approximate percentile rank of 97.5%).

Using the probability distribution, the percentile rank for a ROUGE-L F-score (sc) can be computed by finding the area bounded by the distribution curve to the left of sc. We compare the relative performance of our system and various baselines (mentioned in the following section) through the use of their percentile ranks.

5.5.1 Baselines

The baselines we use for comparison are as follows:

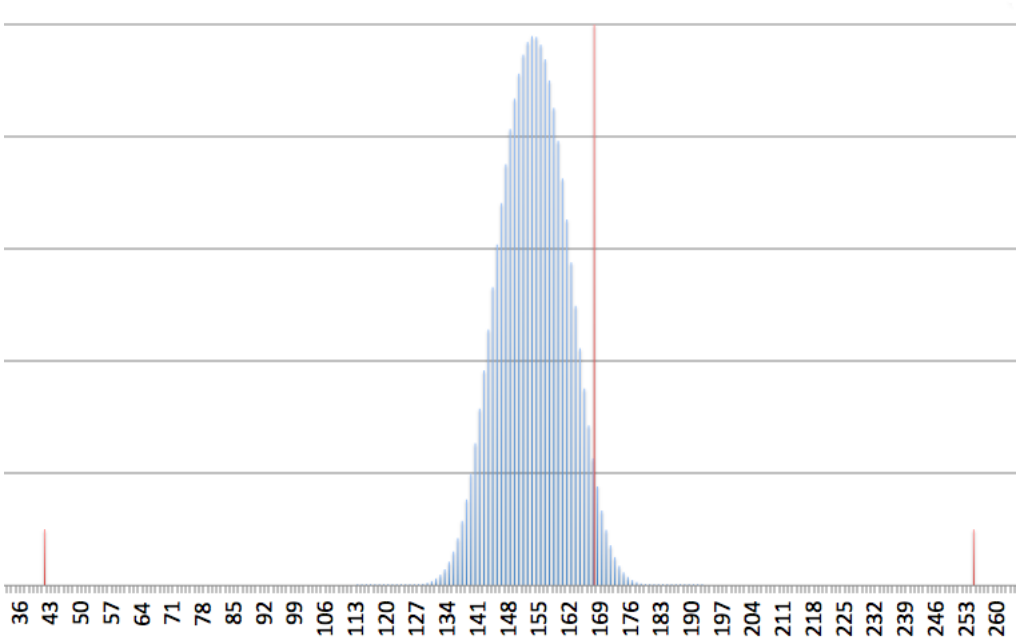


Figure 5.6: The normalised histogram of ROUGE-L F-scores for all abstracts belonging to R_{EVAL}.

Last three sentences.
The last sentences in a medical abstract usually present conclusions, and this has been used as a baseline for summarisation tasks in this domain before [Lin and Demner-Fushman, 2007].

Last three PIBOSO *outcome* sentences.
This is comparable to the summarisation component used by Lin and Demner-Fushman [2007]. In our approach, there can be more than three conclusion sentences. Hence, we use the last three[14]. If there are less than three outcome sentences, all outcome sentences are chosen along with the last occurring non-outcome sentences.

Random.
Three sentences are randomly selected from each abstract.

First three sentences.
This is the baseline used in summarisation for various other domains, the most important being the news domain. Although this baseline is unlikely to perform well in this domain, we are interested in assessing its performance relative to our system and other baselines.

[14]For purely empirical reasons: we have compared this baseline against one that randomly chooses outcome sentences. There is no significant difference in scores.

All Outcomes.
All PIBOSO outcome sentences are chosen as the summary irrespective of the number of sentences. There is no length restriction on the generated summary either.

Sentence position independent.
This is a similar approach to our system, but applying the same statistics for all target sentences. The relative position is used as the score instead of ($RP_{s_{ijkx}}$). This gives a higher score for later sentences. There is no Position Dependent PIBOSO Score. Scores related to question types are also not considered.

Naïve Bayes.
For this baseline, a Naïve Bayes classifier is trained to perform target-sentence-specific extractive summarisation. Naïve Bayes classifiers are commonly used as baselines in standard classification tasks, and therefore we wanted to compare our system with one that models the summarisation task as a sentence classification problem and applies Naïve Bayes to solve it. A separate classifier is trained for each target sentence using the abstracts in R_{TRAIN}. The features used for each sentence are: relative position, length, PIBOSO classification, cosine similarity with question, all the question types for the associated question and the UMLS semantic types present in that sentence. The summarisation task is modelled as a binary text classification problem. All the sentences are divided into two classes: *selected* (meaning that the sentence is in the summary) and *not selected* (meaning that the sentence is not in the summary).

When training the classifier for the first sentence selection ($tn = 1$), for each abstract, only the sentence that is the first sentence in the three best sentences (S_{BEST}) for that abstract is given the *selected* tag. All other sentences in the abstract are given the *not selected* tag and the classifier is trained on this data. Once the classifier has been trained (for the first sentence selection), all the sentences in (R_{EVAL}) are classified using this classifier. Since multiple or no sentences in a single abstract may be classified as *selected*, we do not use the classifications of this classifier. Instead, for each abstract, the sentence with the highest probability of being *selected* is chosen, irrespective of its tag. Ties among sentences with identical probabilities are resolved by choosing the one appearing latest in the document.

Identical approaches are used for the second and the last target sentence. For the latter classifiers, the sentences already selected by previous classifiers are ignored.

SumBasic.
This is a simple summariser [Nenkova and Passonneau, 2005] that performs extractive summarisation by rewarding sentences that contain more frequently occurring non-stop words. Word-frequency based summarisation approaches were very common primarily in early summarisation research. Although we use various frequency distributions in our summarisation task, we do not

actually reward sentences containing *salient terms* with high document frequencies. Therefore, we are interested in comparing our system against one which uses frequency based measures. In SumBasic, each sentence is assigned a score based on the number of high frequency words it contains, and the top scoring sentence is selected.

$$Score(S) = \sum_{w \in S} \frac{1}{|S|} P_D(w) \tag{5.17}$$

where $P_D(w)$ is the observed unigram probabilities obtained from the document D. The summary is progressively built by adding the highest scoring sentence according to the above equation. In order to discourage redundancy, the probabilities of the words in the selected sentence are updated: $P_D^{new} = P_D^{old}(w)$. Despite its simplicity, it has been shown to yield good results at the DUC 2006 evaluation[15].

FastSum (modified).
FastSum [Schilder and Kondadadi, 2008] is a fast query-focused multi-document summariser based solely on word-frequency features of topics, documents, and clusters. This summarisation system is designed for multi-document summarisation, and so we modify some of the features to apply it to single-document summarisation. Each sentence is ranked using a linear equation of scores based on various frequency measures. The weights for the features are learned by a regression SVM. Since we apply the system to single-document summarisation, unlike the original paper describing that work, we do not perform QR decomposition with pivoting for reducing redundancy.

MEAD.
MEAD [Radev et al., 2004] is a platform (not a summarisation system) for multi-document summarisation that provides various functionalities, including the generation of query-focused summaries. We customised this platform according to our needs to make it suitable for our domain. Since MEAD was initially designed for news summarisation, its *position* feature heavily rewards sentences that appear earlier in the document. To make it suitable for our domain, we reverse this default feature to make later sentences more likely to be chosen. We also relax the sentence length related constraint used in the default version of MEAD. We add scores based on sentence-query similarity, and tune several other parameters provided by the platform. We also add additional scores for each sentence including the PIBOSO and query related scores. Thus, while this customisation of MEAD does not resemble any existing summarisation system, it is a baseline that shares most of the features as our system (except for a target-sentence-specific score for the relative position and a length associated score), and therefore its performance is likely to be similar to ours.

[15] http://www-nlpir.nist.gov/projects/duc/guidelines/2006.html. Accessed on 26th May, 2014.

System	F-Score	95% CI	Percentile Rank (%)
Last Three	0.15482	0.151 - 0.158	55.9
Last Three Outcome	0.15920	0.155 - 0.163	74.2
Random	0.15251	0.149 - 0.156	46.1
First Three	0.13994	0.136 - 0.143	36.9
All Outcomes	0.15936	0.155 - 0.164	74.2
Position Independent	0.16019	0.157 - 0.164	78.1
Naïve Bayes	0.15551	0.152 - 0.159	55.9
SumBasic	0.15818	0.155 - 0.162	69.9
FastSum (modified)	0.15769	0.154 - 0.161	69.9
MEAD	0.16332	0.160 - 0.167	85.4
QSpec (without query types)	0.16627	0.162 - 0.170	94.3
QSpec	**0.16780**	0.164 - 0.172	96.8
QSpec (regression)	0.16479	0.161 - 0.169	92.5

Table 5.5: ROUGE F-scores, 95% confidence intervals and percentile ranks for our system and several baselines.

5.5.2 Results

Table 5.5 presents the ROUGE-L F-scores for our system and the baselines, the 95% confidence intervals (CI) for the F-scores as reported by ROUGE, and the percentile rank for each score. Note that the ROUGE-L F-scores are generally much lower than F-scores for other tasks (*e.g.*, text classification) because of the complexities of the summarisation tasks. In the table, QSpec represents our system in its full implementation. QSpec outperforms all systems with a percentile rank of 96.8%. It is, however, closely followed by a version of QSpec that does not use query type information. Learning the weights via regression results in slight degradation of performance, but is still better than the other systems. None of these three versions of our system are statistically significantly better than each other. This shows that the use of statistics derived from our specialised corpus adds a massive advantage for our summarisation system. This, in turn, justifies the use of specialised corpora for such domain specific summarisation tasks.

As expected, random three sentence summaries produce ROUGE scores that are very close to the centre of the distribution shown in Figure 5.6. Our customised MEAD summarisation system is the only system that manages to have a score that does not lie below the lower CI limit of the best performing system. This suggests that this summarisation platform has the potential of being customised to restricted domains such as ours. Following MEAD, the next best performing baseline is our Position Independent system. This is closely followed by the *Outcome*-based systems, one of which is our implementation of the system proposed by Lin and Demner-Fushman [2007]. Naïve Bayes, SumBasic and FastSum do not perform particularly well,

and the baseline of the first three sentences has the worst performance. The poor performances of SumBasic and FastSum indicate that word-frequency based approaches are perhaps not suited for this domain.

Since we use 1,000 bins when generating the document histograms, our evaluation mechanism does not differentiate between ROUGE scores beyond the third decimal place. Therefore, some baselines have identical percentile ranks despite their ROUGE-L F-scores being different. Using smaller bin sizes (*i.e.*, larger number of bins) would give more accurate measures[16]. For our analysis, however, we find this level of granularity sufficient. Our approach is also fast, since it does not involve any time-consuming computation such as parsing. This is an important property of our system and is in line with our intent of reducing the time needs of evidence-based medicine practice. Figure 5.7 presents a sample extractive summary generated by QSpec along with the associated human authored summary. A random sample of QSpec summaries can be found in Appendix C.

Question: Are there big differences among beta-blockers in treating essential hypertension?

QSpec summary: Because the pathophysiology of hypertension differs in older and younger patients, we designed this meta-analysis to clarify the efficacy of beta-blockers in different age groups. In placebo-controlled trials, beta-blockers reduced major cardiovascular outcomes in younger patients (risk ratio [RR] 0.86, 95% confidence interval [CI] 0.74-0.99, based on 794 events in 19 414 patients) but not in older patients (RR 0.89, 95% CI 0.75-1.05, based on 1115 events in 8019 patients). Beta-blockers should not be considered first-line therapy for older hypertensive patients without another indication for these agents; however, in younger patients beta-blockers are associated with a significant reduction in cardiovascular morbidity and mortality.

Human authored summary: A meta-analysis found that beta-blocker therapy in younger patients (less than 60 years of age) is associated with a significant reduction in cardiovascular morbidity and mortality. Researchers used data from 145,811 participants in 21 hypertension trials, beta-blockers reduced major cardiovascular outcomes in younger patients (risk ratio=0.86; 95% CI, 0.74-0.99) but not in older patients (risk ratio=0.89; 95% CI, 0.75-1.05).

Figure 5.7: A sample 3-sentence, query-focused, extractive summary generated by QSpec.

Analysis of Individual Feature Contribution

To end this section, we provide a brief analysis of the individual features used for scoring and their importance in our summarisation system. In order to assess the contribution of each feature

[16]The algorithm for convolving *pdf*s has a complexity of $O(n^3)$, and increasing the number of bins by a factor of 10 (*i.e.*, to incorporate another decimal place) will increase the running time of the program by a factor of 1,000.

for the final score, we performed two simple ablation experiments – (i) performing sentence scoring using single features only, and (ii) performing sentence scoring by leaving out one feature. Tables 5.6 and 5.7 present the ROUGE-L F-scores for both these experiments (i and ii respectively) on the training set as well as the test set. From table 5.6, it can be observed that all the single features scores indicate significant improvements over the score that is obtained using no features (*i.e.*, summaries are the first three sentences). All these improvements are statistically significant, as reported by ROUGE. The scores for the test set are slightly higher compared to those in the training set[17], and the best performing feature for each of the two sets is shown in boldface. Importantly also, none of the single feature scores are better than the score obtained by the combination of features. The last row in table 5.6 presents the results when only target-sentence-specific features are used (*i.e.*, sentence position, PDPS, and MMR). It can be seen that the combination of the three target-sentence-specific scores performs better than each of the individual feature scores, but not as good as all the scores combined. The same is true for the leave-one-out scores shown in table 5.7, although the scores are much higher than the individual feature scores. None of the scores in table 5.7 are statistically significantly lower than the best score obtained by combining these features, which indicates that the final score is not biased by the influence of a single score. The lowest ROUGE-L F-scores are presented in bold-face, and these show the largest drops in the ROUGE score when those features are removed. From the two tables it can be seen that MMR and the question specific semantic type features appear to be the most useful features. It must be mentioned here that the differences in performance numbers between the best single feature system, the best leave-one-out feature system, and the best final system are quite small. This happens because of the distribution of the ROUGE scores for all possible summary combinations (*i.e.*, all summary scores tend to lie around the mean score). Thus, small differences in scores are likely to represent important differences in performance. Manual analysis to compare the correlation between changes in summary scores and changes in performance need to be performed for detailed evaluation of the features. However, such a manual evaluation is outside the scope of this thesis.

5.6 Summary So Far

In this chapter, we addressed the first step of the two-step evidence-based summary generation process. This first step involves generating single-document, query-focused summaries. We modelled this first step as a query-focused, extractive summarisation problem with sentences being the target extracts. Our target was to implement a summarisation technique that extracts

[17] It is interesting to observe that the test set scores are greater than the training set scores. Our investigations revealed that this is because the training set ROUGE scores' *pdf* is slightly shifted to the left compared to the test set. As such, despite the slightly lower ROUGE scores for the training set, their percentile ranks are comparable to the percentile ranks of the test set scores.

Feature	Training Set	Test Set
None	0.13891	0.13994
Random	0.15251	0.15316
Relative Position	0.15301	0.15372
Length	0.15746	0.16066
Position Independent PIBOSO	0.15791	0.16204
Position Dependent PIBOSO	0.15497	0.15532
Question Specific Semantic Type	0.16091	**0.16397**
Question Specific Association	0.15657	0.15797
MMR and Similarity ($\lambda = 0.0$)	0.14882	0.14971
MMR and Similarity ($\lambda = 0.5$)	0.15698	0.16130
MMR and Similarity ($\lambda = 1.0$)	**0.16127**	0.16353
Target-sentence-specific features only	0.16281	0.16551

Table 5.6: Single feature scores for the training and evaluation sets.

Feature left out	Training Set	Test Set
Relative Position	0.16321	0.16643
Length	0.16476	0.16728
Position Independent PIBOSO	0.16365	0.16718
Position Dependent PIBOSO	0.16394	0.16693
Question Specific Semantic Type	0.16252	**0.16485**
Question Specific Association	0.16495	0.16730
MMR	**0.16172**	0.16546

Table 5.7: Leave-one-out scores for the training and evaluation sets.

three sentences from source documents, based on the information needs of a given query, such that the extracted sentences closely resemble human-authored summaries.

We used a corpus that is specifically designed for evidence-based summarisation for our analysis and summarisation approach. We divided the corpus into two parts — one for obtaining statistics and the other for evaluation. From the source documents in the training set, we generated the best three sentence summaries. We then used these best sentences and also the human-generated summaries associated with each source document to compute various statistics which we used for summarisation. We used features such as relative sentence positions, sentences lengths, the PIBOSO classifications of sentences, similarities between sentences and the associated queries, the semantic types present in sentences, and the semantic associations between sentences and the associated queries. We showed that using carefully extracted statistics from a specialised corpus significantly improves summarisation performance. We applied a strategy, which we call target-sentence-specific summarisation. Using this strategy, we applied different statistics for different target summary sentences. We also modified the MMR approach to enable concept matches, and significantly utilised available domain knowledge . Furthermore, we showed that customising the summarisation technique to the type of question improves summarisation performance.

We used the automatic summary evaluation tool ROUGE to evaluate our extractive summaries relative to the human-generated summaries. We compared our summarisation system to various established baselines for this domain using a percentile-rank based approach. The best ROUGE-L F-score obtained by our system has a percentile rank of 96.8%, which is a statistically significant improvement over the best performing baseline system.

Based on the findings presented in this chapter, we make several conclusions. First of all, our results clearly show that an extractive summarisation approach such as ours can be effectively used for selecting informative content from source documents. The contents may be processed later for multi-document summarisation and generation of bottom-line answers. Secondly, for a domain as complex as the medical domain, it is crucial to incorporate domain knowledge. We incorporated domain knowledge into our system in various ways — either directly (*e.g.*, through the use of MetaMap), or indirectly (*e.g.*, through the classification of sentences into PIBOSO elements) — and showed its importance in the summarisation task. Thirdly, the use of target-sentence-specific summarisation can improve the performance of a summarisation system by allowing the selection of diverse content. This approach is easily portable, and it will be interesting to see how it performs in other domains. Finally, and perhaps most importantly, we conclude that the use of specialised corpora is vital for such domain-specific summarisation tasks.

Our summarisation approach is extractive, and we do not take into consideration factors such as summary coherence. Instead, our focus is to select informative sentences that can be used to

generate bottom-line answers. Chapter 6 describes our research work on generating bottom-line summaries by utilising the single-document summaries that we generate automatically.

6 Towards Multi-document Summarisation

6.1 Introduction

As we have noted, our model to automatically generate evidence-based answers to clinical queries involves two processes: (i) assessing the quality of evidence and expressing it through the use of evidence grades, and (ii) summarising the contents of medical documents to generate short, evidence-based answers. In Chapter 4, we explored the problem of automatically generating of evidence grades. Using a supervised machine learning model, we first analysed various features and their usefulness in predicting evidence grades, and we proposed an approach for sequentially combining high precision classifiers to minimise errors when predicting evidence grades.

In Chapter 5, we focused our attention to automatic text summarisation and attempted to perform query-focused, extractive, single-document text summarisation. By studying the human-authored, single-document summaries in our corpus, we analysed the contents of evidence-based summaries. Our analysis revealed that the summaries tend to contain diverse information from the source articles, including some background and outcome information. Employing a simple extractive summarisation model based on the Edmundsonian Paradigm, we explored possible approaches by which distinct text nuggets containing the most pertinent information can be extracted. We showed that the performance of summarisation systems in a specialised domain such as ours can significantly improve if statistics are derived from a specialised corpus, and through the incorporation of various domain-specific information. In our approach, we incorporated domain-specific information in various ways, including the use of semantic similarity with lexical similarity, the incorporation of semantic types, semantic associations between domain-specific concepts, sentence types, and question types. We also applied target-sentence-specific statistics, an approach by which the same sentence may get a different score depending on the target summary sentence number. We evaluated our approach automatically using percentile-ranks of ROUGE scores and showed that our system produces content-rich summaries and performs

significantly better than various baseline and benchmark approaches.

In this chapter, we follow on from the research discussed in Chapter 5 and examine the final component of our summarisation task: the generation of bottom-line summaries. The bottom-line summaries are generally produced via the synthesis of information from multiple documents, and in our work, we attempt to produce them from the query-focused single-document summaries produced by the approach described in Chapter 5. In particular, this chapter focuses on the exploration of the following two aspects of bottom-line summary generation:

i the possibility of generating bottom-line summaries from the single-document summaries, instead of full source abstracts; and

ii possible approaches that can be applied for the generation of bottom-line summaries from the single-document extracts.

It must be mentioned that our intent is not to produce complete multi-document summaries, but to explore possible approaches that can be applied for the task. We first perform some experiments to assess how much useful information from source abstracts is retained by various types of summaries. The experiments let us quantify the limits of a bottom-line summary generator that relies on single-document summaries as inputs. The experiments also enable us to analyse how much information is lost during the single-document summary generation process. We introduce several variants of *coverage scores*, and we use them to analyse the extent to which the contents of bottom-line summaries are contained in source texts of various granularities, including full abstracts, human-authored summaries, and system generated summaries. Our analyses show that good summaries tend to contain important information associated with queries, but in a much compressed manner compared to the full abstracts. More specifically, we apply a variety of measures in our *coverage analysis* to show that content-rich summaries contain lower amounts of noisy information compared to full source abstracts. We argue that it may not only be possible but also useful to use the single-document summaries instead of full abstracts for the generation of bottom-line recommendations.

Following our *coverage analysis* work, we investigate the applicability of two genres of abstractive summarisation approaches for the task of bottom-line summary generation from single-document summaries. Our investigations suggest that redundancy-reliant approaches such as *Sentence Fusion* [Barzilay and McKeown, 2005] may not be ideal for multi-document summarisation in our domain, because of the low degrees of similarities/redundancies among sentences in this domain. We also explore automatic sentence level, context-sensitive polarity classification as a possible approach for the generation of bottom-line summaries. We show that automatic polarity classification, representing the class of sentiment analysis/opinion summarisation approaches

that use supervised machine learning, is a promising approach for the generation of bottom-line summaries. We select a subset of summary sentences from the corpus, manually annotate the contexts and context-sensitive polarities, and, using a combination of context-sensitive and context-free features, show that high accuracies can be obtained using a supervised classification model.

The rest of this chapter is divided into two broad sections. In Section 6.2, we describe our coverage analysis work and show that our single-document summaries provide relatively good coverage of the bottom-line summaries, compared to the full abstracts. In Section 6.3, we analyse the applicability of sentence fusion and sentence level polarity classification as two candidate approaches for the generation of abstractive, bottom-line, evidence-based summaries, using automatically generated single-document summaries as input. In Section 6.4, we conclude the chapter and provide a discussion of the possible future steps required to produce fully automatic, abstractive, bottom-line summaries to answer questions with evidence.

6.2 Coverage Analysis

As we explained earlier in this thesis, to generate evidence-based answers to clinical queries, practitioners utilise the best available clinical evidence along with their own expertise. However, there is no measure of the extent to which human experts incorporate their own expertise in the evidence-based answer generation process. Our intent is to use the query-focused, single-document summarisation approach described in Chapter 5, as a content selection step for the bottom-line summary generation process. Although we have shown that our summarisation approach in Chapter 5 generates high quality summaries, we have no measure of how much information is lost during summary generation. The experiments described in this section serve two purposes:

1. They help make *estimations* about the extent to which human experts rely on the information presented in clinical texts to generate evidence-based answers.
2. They provide evidence supporting the use of single-document summaries to generate bottom-line answers. Thus, our overall summarisation approach can be viewed as a two-step process, with the single-document summarisation approach being the content-selection step.

To perform this analysis, we compare bottom-line evidence-based summaries (a_i) to the associated set of source texts (D_i), and the associated set of human-generated, query-focused, single-document summaries (L_i) of the source texts. This allows us to determine if good single-document summaries contain sufficient content, from source texts, to be used for the generation of

multi-document, bottom-line summaries. The human-authored summaries, in this case, represent gold abstractive summaries of the source texts. We also study the single-document extractive summaries generated by various summarisation systems, and compare their performance relative to source texts and human-generated summaries. In addition to this analysis, we attempt to make *estimations* about the extent to which the core contents of the bottom-line summaries come from the source texts. Such an analysis is of paramount importance in this domain because, if only a small proportion of the bottom-line summaries contain information from the source articles, we can assume that the summaries are almost entirely generated from specialised human knowledge, making it impossible to perform text-to-text summarisation automatically in this domain without intensive use of domain-specific knowledge.

6.2.1 Coverage Scores

Our first analytical experiments attempt to estimate the extent to which information in the set of bottom-line summaries, A, is contained in the source document abstracts, D_a, associated with each summary (a). This gives us a measure of the extent to which extra information are added to the final summaries by the authors of the articles from which the corpus has been built. For this, we define a set of scores, which we call *coverage scores*. The greater the score, the better the bottom-line summary coverage. Consider a bottom-line summary a, which contains a set of m terms, and the associated source documents, D_a. The first variant of the coverage scores that we use is a term-based measure and is given by the following equation:

$$Coverage(a, D_a) = \frac{|a \cap D_a|}{m} \tag{6.1}$$

where $|a \cap D_a|$ represents the number of terms common to a bottom-line summary (a) and the associated source texts (D_a). We first preprocess the text by removing stop words and punctuations, lowercasing all terms and stemming the terms using the Porter stemmer [Porter, 1980]. Term tokenisation is performed using the word tokeniser of the NLTK[1] toolbox. Such a term-level coverage measurement, however, often fails to identify matches in the case of medical concepts that may be represented by multiple distinct terms. An example of this is the term *high blood pressure*. In our corpus, this term has various other representations including *hypertension* and *hbp*.

Figure 6.1 illustrates term-level coverage with an example of a bottom-line summary and one of the associated detailed justifications (human-authored summary). The common terms between

[1] nltk.org

Question:
What medications are effective for treating symptoms of premenstrual syndrome PMS?

Bottom-line answer:
Selective serotonin reuptake inhibitors (SSRIs) and some other antidepressants are more effective, but are also more costly and more likely to cause side effects or treatment dropout.

Summary:
Among SSRIs, fluoxetine 20 mg/d is well-studied and effective. Gonadotropin-releasing hormone agonists may be effective, but troublesome anti-estrogenic side effects limit their utility. Estrogen and progesterone add-back therapy to counter side effects further complicates this approach. The gonadotropin inhibitor danazol has a high treatment dropout rate at higher doses 200-400 mg/d continuously, but can be effective in individuals who are able to tolerate it.

Figure 6.1: Illustration of bottom-line summary terms covered in our term-level coverage computation.

the two summaries are shown in red. The bottom-line summary contains a total of 23 unique terms, of which 10 are covered. Thus, using our equation above, this would give a coverage score of 0.43. Note that, for this particular bottom-line summary, there are additional associated human-authored single-documents summaries, which we do not show here for simplicity. We do not remove stop words from the example for the same reason. The example also shows that, although the terms 'Selective serotonin reuptake inhibitors' should be covered because they represent the same concept as SSRI, our term-level coverage approach fails to detect this.

Incorporation of CUIs and Semantic Types

To address the issue of same concepts having distinct lexical representations, we identify the semantic types and Concept Unique Identifiers (CUI) of all the terms in the corpus and incorporate this information in our coverage computation. Using CUIs in the computation reduces the dependence on direct string matching because distinct terms representing the same medical concept have the same CUI. For example, all the different variants of the term *high blood pressure* have the same CUI (C0020538). However, it is also possible for terms with different CUIs to have the same underlying meaning in our corpus. For example, the terms [African] women (CUI: C0043210) and African American (CUI:C008575) have different CUIs but have been used to represent the same population group. The two terms have the same (UMLS) semantic type: *popg* (*population group*), and this information may be used to match the two terms in our experiments.

We use the MetaMap[2] tool to automatically identify the CUIs and semantic types for all the text in our corpus.

We introduce two variants of the coverage scores in addition to the term-level coverage computation. In our first variation, we use individual terms and CUIs; and in the second variation we use terms, CUIs and semantic types. We apply a sequence of functions that, given a and D_a, along with the CUIs and semantic types of the terms in a and D_a, compute $a \cap D_a$ utilising all the available information (*i.e.*, terms, CUIs, semantic types). Term-based matching is first performed, and the terms in a that are exactly matched by terms in D_a are collected. Next, for the unmatched terms in a, CUI matching is performed with the CUIs of D_a. This ensures that different lexical versions of the same concept are detected correctly. All the matched terms are added to the covered terms collection. In our first variant, this value is used for $|a \cap D_a|$ in equation 6.1. For the second variant, for terms that are still uncovered after CUI matching, semantic type matching is performed, and the terms in a with matching semantic types are added to the covered terms collection before computing the coverage score.

A problem with the use of semantic types in coverage score computation is that they are too generic and produce many incorrect matches. For example, the terms *pneumonia* and *heart failure* are two completely distinct concepts but have the same semantic type (*disease or syndrome (dsyn)*). The use of semantic types, therefore, leads to incorrect matches, resulting in high coverage scores. We still use semantic types along with terms and CUIs in our experiments because their coverage scores give an idea of the coverage upper limits.

Figure 6.2 illustrates how the CUI and semantic type variants of the coverage scores are computed. The figure shows the elements remaining after term-level coverage along with their CUIs. It can be seen that the two lexical representations of SSRIs have the same CUI (C0162758) and therefore, the terms representing this concept are added to the list of covered elements. After the CUI-based matching, a total of 14 terms in the bottom-line summary are covered, and so the new coverage score is 0.61. Finally, the figure also shows that two more terms are covered when semantic type matching is performed (shown in red). This results in a total of 16 elements to be covered, giving a coverage score of 0.70. The figure also illustrates the problem associated with the semantic type variant of the coverage score; although danazol is not an antidepressant, it belongs to the same semantic category as antidepressants (*i.e.*, *pharmacological substance (phsu)*), resulting in a false positive.

[2]http://metamap.nlm.nih.gov/. Accessed on 26th May, 2014.

Text elements remaining after performing term-level coverage along with CUIs:
Selective serotonin reuptake inhibitors (C0162758), some, other, antidepressants (C0003289), more, also, costly, likely (C0332148), cause (C1524003), or

Matching CUI(s) from human-authored summary:
SSRI (C0162758)

Text elements remaining after performing CUI-level coverage along with semantic types:
some, other, antidepressants (phsu), more, also, costly, likely (qlco), cause (cnce), or

Matching semantic type(s) from human-authored summary:
danazol (phsu), effective (qlco)

Final uncovered elements:
some, other, more, also, costly, cause, or

Figure 6.2: Illustration of elements covered in our CUI and semantic type variants of coverage computation.

Concept Coverage

In an attempt to reduce the number of non-medical terms in our coverage score computation, we introduce a fourth variant to our coverage scores which we call *Concept Coverage* (CC). We noticed that, often, non-medical terms such as entities and numbers are the primary causes of mismatch amongst different terms. This coverage score only takes into account the concepts (CUIs) in a and D_a. Referring to equation 6.1, m in this case represents the number of unique CUIs in a, while $|a \cap D_a|$ is computed as a combination of direct CUI matches and similarity measures among unmatched CUIs. That is, besides considering direct matches between CUIs, we also consider *similarities* among concepts when performing this calculation. This is important because often bottom-line summaries contain generic terms representing the more specific concepts in the source texts (*e.g.*, the generic term *anti-depressant* in the bottom-line summary to represent *paroxetine*, *amitriptyline* and so on). The concept similarity between two concepts gives a measure of their *semantic relatedness* or how *close* two concepts are within a specific domain or ontology [Budanitsky and Hirst, 2006].

In our concept coverage computation, each CUI in a receives a score of 1.0 if it has an exact match with the CUIs in D_a. For each unmatched CUI in a, its concept similarity value with each unmatched concept in D_a is computed, and the *maximum similarity* value is chosen as the score

for that concept. To compute the similarity between two concepts, we use the similarity measure proposed by Jiang and Conrath [1997]. The authors apply a corpus-based method that works in conjunction with lexical taxonomies to calculate semantic similarities between terms, and the approach has been shown to agree well with human judgements. We use the implementation provided by McInnes et al. [2009], and scale the scores so that they fall within the range [0.0,1.0), with 0.0 indicating no match and 0.99 representing near perfect match (theoretically). The direct match scores or *maximum similarity* scores of each CUI in a are added and divided by m to give the final concept coverage score.

Comparison of Coverage Scores

Our intent is to determine the extent to which the contents of the bottom-line summaries in the corpus are contained in source texts of different granularities. This gives us an estimate of the information loss that occurs when source text is compressed by various compression factors. More specifically, in our experiments, a (in equation 6.1) is always the bottom-line summary, while for D_a, we use the following variants:

i all the text from all the article abstracts associated with a (FullAbs),

ii all the text from all the human-authored, single-document summaries (from L) (HS),

iii all the text from all the single-document, three sentence extractive summaries, produced by our system (QSpec), associated with a,

iv all the text from all the *ideal* three sentence extractive summaries associated with a (IdealSum), and

v all the text from random three sentence extractive single-document summaries associated with a (Random).

The IdealSum summaries are three sentence, single-document, extractive summaries that have the highest ROUGE-L *F-scores* when compared with the human-generated single-document summaries (l)[3]. Using these five different sets enables us to estimate the degradation, if any, in coverage scores as the source text is compressed.

For each data set, we also compute their ROUGE-1 recall scores (after stemming and stop word removal) with the bottom-line summaries, and compare these scores. This enables us to compare the coverage of these data sets using another metric, in addition to the coverage scores.

[3]These summaries were produced by generating all three sentence combinations for each source text, and then computing the ROUGE-L F-score for each combination. Further details about the generation of these summaries were provided in Chapter 5.

System	T	T & C	T, C & ST	CC
FullAbs	0.596	0.643	0.782	0.659
QSpec	0.502	0.546	0.683	..
HS	0.595	0.630	0.737	0.644
IdealSum	0.468	0.511	0.654	..
Random	0.403	0.451	0.594	..

Table 6.1: Coverage scores for the five data sets with the bottom-line summaries. T = Terms, C = CUIs, ST = Semantic Types, and CC = Concept Coverage.

6.2.2 Coverage Analysis Results and Evaluation

Table 6.1 shows that, when terms and CUIs are used, the source texts cover approximately 65% of the summary texts, and incorporating semantic types takes the coverage score close to 80%. The concept coverage scores are similar to the term and CUI (T & C) overlap scores. The analysis of the *uncovered* components reveals a number of reasons behind coverage mismatches. First of all, authors often prefer using generalised medical terms in the bottom-line summaries, while the source texts contain more specific terms (*e.g.*, *antibiotics vs. penicillin*). Incorporating semantic types ensures coverage in such cases, but also leads to false matches. Secondly, MetaMap does not have perfect word sense disambiguation accuracy [Plaza et al., 2011b] and often fails to disambiguate terms correctly, causing variants of the same term to have different CUIs, and often different semantic types. Thirdly, a large portion of the uncovered components consists of text that improves the qualitative aspects of the summaries and do not represent important content. In other words, most of the mismatches are due to *generic* content, rather than *specific* content. In an analysis by Louis and Nenkova [2011] on news text, they showed that summaries, particularly human-authored summaries, tend to contain significant amounts of *generic* information along with *specific* information. If we consider the medical concepts in the summaries as *specific* contents and the non-medical terms as *generic* contents, then it is not surprising that the texts of all granularities contain significant amounts of generic information, which may be added or lost during summarisation.

Table 6.1 reveals that the human-generated single-document summaries (HS) have almost identical coverage scores to full source articles[4]. Figure 6.3 shows the distributions of the concept coverage scores for the two sets, and it can be seen that the distributions are very similar. The coverage scores obtained by the two summarisation systems (IdealSum and QSpec) also have high coverage scores compared to the Random summaries. Interestingly, the QSpec summaries actually produce slightly better coverage than the IdealSum summaries. Manual analysis shows that this can be attributed to two reasons: (i) the three sentence summaries of our system tend to be slightly longer

[4]We only compute the concept coverage scores for the FullAbs and HS sets because of the extremely long running time of the similarity measurement algorithm.

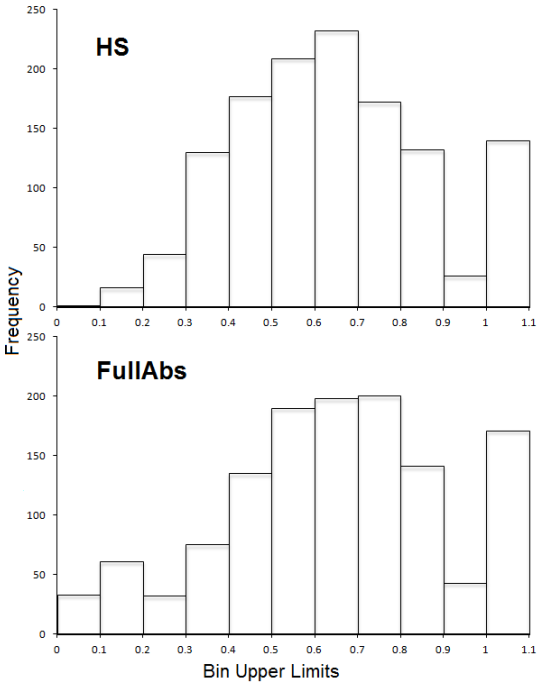


Figure 6.3: Distributions for concept coverage scores.

and, (ii) they tend to be richer in terms of domain specific content. Random single-document summaries fail to produce up to 60% coverage. Importantly, the table suggests that a significant component of the bottom-line summaries does come from the information present in the medical texts, thus making it theoretically possible to perform automatic summarisation to generate the bottom-line summaries.

Table 6.2 shows that the ROUGE-1 recall scores are also very similar for the HS and FullAbs data sets and lie within each other's 95% confidence intervals, indicating that there is no statistically significant difference between the contents of the HS and FullAbs sets. The coverage scores and the ROUGE-1 scores illustrate that the human-authored summaries, which we consider to be the *ideal* abstractive summaries, contain similar amounts of relevant information as the source articles themselves.

To verify if the differences in the coverage scores between the HS and FullAbs sets are statistically significant, we perform statistical significance tests for the two pairs of coverage scores. Due to the paired nature of the data, we perform the Wilcoxon rank sum test with the null hypothesis that the coverage scores for the two sets are the same ($\mu_0 = 0$). Table 6.3 shows the z and p-values for the tests performed for the term, term and CUI and concept coverage scores for the HS and FullAbs sets. In all cases $p > 0.05$, meaning that we cannot reject the null hypothesis. Therefore, the difference in the two sets of coverage scores is not statistically significant. This

System	Recall	95% CI	Compression Factor
FullAbs	0.418	0.40 - 0.44	0.05
HS	0.405	0.39 - 0.42	0.15
QSpec	0.318	0.30 - 0.34	0.26
IdealSum	0.284	0.27 - 0.30	0.20
Random	0.229	0.21 - 0.24	0.21

Table 6.2: ROUGE-1 recall scores and 95% confidence intervals for the five data sets with the bottom-line summaries.

	T	T & C	CC
z	-1.5	-1.27	-1.33
p-value (2-tail)	0.13	0.20	0.16

Table 6.3: z and p-values for Wilcoxon rank sum tests.

adds evidence to the hypothesis that good single-document summaries may contain sufficient content for bottom-line summary generation. This, in turn, supports the proposition that the generation of bottom-line summaries *may be* modelled as a two step process, in which the first step involves summarising individual documents, based on the information needs of queries, and the second step synthesises information from the individual summaries.

Table 6.2 also shows the compression factors (CF) for each type of source texts. The CFs show the relative compression rates required for the various source texts to generate the bottom-line summaries. The compression factor is computed as: $|B|/|A|$, where $|B|$ represents bottom-line summary length and $|A|$ the source text length. This means that the higher the value, the easier should the summarisation technique be because of the lower amount of source text compression required. It can be seen that generating bottom-line summaries from original source texts requires approximately 5 times more compression compared to the generation from single-document summaries, suggesting that the single-document summaries contain important information from the source texts in a much compressed manner. Thus, for a summarisation system that focuses on generating bottom-line summaries, it is perhaps better to use single-document summaries as input rather than whole source texts, as the information in the source texts is generally very noisy. Considering the balance between coverage scores and compression factors of IdealSum and QSpec, such content-rich automatic summaries *may* prove to be preferable inputs for the generation of bottom-line summaries over original texts.

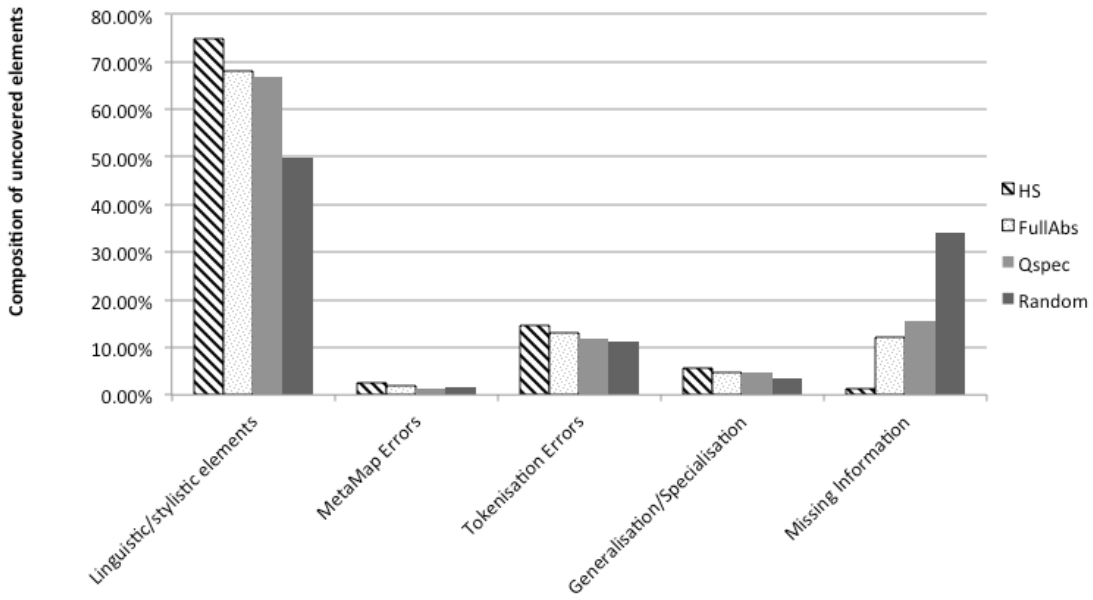


Figure 6.4: Comparison of the composition of uncovered elements between the HS, FullAbs, QSpec and Random sets.

6.2.3 Analysis of Uncovered Elements

Our coverage analysis shows that the human-authored summaries contain the important content from the source abstracts, but in a more compressed form. The analysis supports the hypothesis that the generation of bottom-line summaries can be modelled as a two-step process. However, if a fully automatic two-step approach is to be applied, we cannot rely on human-authored summaries for the first step. Our intent is to use the content-rich summaries generated by QSpec as inputs for the second step of the summarisation process. Since, in terms of coverage, there is a significant difference between the FullAbs and QSpec texts, we manually analyse the *uncovered* elements from the bottom-line summaries to quantitatively verify the type of content contained in the *uncovered* elements and the causes for *non-coverage*. This analysis is important to assess the applicability of our summarisation approach: if too much information is missing from the single-document summaries, they cannot be utilised for the multi-document summarisation task.

To perform our analysis, we randomly select 31 records from our corpus and manually inspect each element from the bottom-line summaries that are not covered by the source texts and the various summaries. We manually classify the uncovered elements into five distinct categories: (i) stylistic content, (ii) tokenisation errors, (iii) generalisation/specialisation, (iv) MetaMap errors, and (v) important missing content, as defined below.

HS. We commence this analysis by inspecting the uncovered elements associated with the summaries in the HS set. The bottom-line summaries of the 31 records contain a total of 2,309

elements (tokens) of which 708 (30.7%) are not covered when the Term-plus-CUI based coverage score is used. We categorise 530 (74.86%) of the 708 uncovered elements as stylistic/non-medical content. These include the use of distinct English terms/phrases between the single-document summaries and bottom-line summaries, some of which are synonyms or paraphrases (*i.e.*, containing same or similar information). Examples include: *young people vs. adolescents*, and *benefit vs. improve*. In some cases, the uncovered elements are numbers, or numerical concepts and symbols, that are not presented in a similar way in the single-document summaries (*e.g.*, *3 × vs.* three times). Terms are only added to this category when the ideas or concepts presented by them are also conveyed by the human-authored summaries, but through the use of different lexical terms. For example, a bottom-line summary may have three associated human-authored single-document summaries. Each individual single-document summary may present the outcomes with the same polarity, worded in different ways. Generally, these summaries contain more details than the bottom-line summaries, including information about results and outcomes. The bottom-line summary, however, generally just states the final recommendation (*e.g.*, *drugs x, y, and z are recommended for disease d*). Even if the contents of the single-document summaries represent the same information as the associated bottom-line summary, the exact terms in the bottom-line summary may not be contained in them, leading to non-coverage. We provide examples later on in this section.

Word tokenisation errors account for 103 (14.55%) of the uncovered elements. Although the information represented by these elements is actually covered by the single-document summaries, tokenisation errors cause the automatic coverage computations to fail. We use the default word tokeniser in MetaMap, and the tokeniser often fails in the case of multi-word expressions. Some examples of tokenisation errors include: *sodium/sodium valproate*, *prilocaine-lidocaine*, and *paroxetine-plus-emla*. These tokenisation errors also result in the failure of MetaMap to correctly identify medical concepts and semantic types.

The use of generic medical terms in the bottom-line summaries to represent a number of specific terms in the single-document summaries occur in 40 cases (5.65%), resulting in non-coverage. Examples include: *antibiotic* to represent *amoxicillin* and *penicillin*; *anti-convulsant* for *valproate* and so on. These cases mostly occur when the individual summaries mention specific drug/chemical names, while the associated bottom-line summaries use the generic terms representing the whole class of specific names. These cases, however, are covered in the variant of the coverage score that incorporates the medical semantic types. Also, these non-coverage cases do not actually represent information that is missing from the single-document summaries.

For 19 cases (2.68%), non-coverage was caused by MetaMap disambiguation error. These are cases when there are no tokenisation errors, but MetaMap fails to identify medical concepts correctly. For example, in two cases *hbp* is identified as a protein instead of a terminological

variant for *hypertension* or *high blood pressure*. Another example is the failure to recognise *tb* as the same concept as *tuberculosis*.

Only 10 of the 708 uncovered elements can be considered to be useful information missing from the single-document summaries. These cases include drug names (*e.g.*, *calcium tablets*) or other useful information such as diagnostic procedures (*e.g.*, *x-ray*). This represents only 1.41% of all the uncovered elements in that analysis set. We can thus conclude that good single-document summaries (in this case, human-authored ones) do contain sufficient information from the source documents and therefore may be used for the generation of bottom-line summaries via synthesis of information.

FullAbs. We also analyse the uncovered elements from the full abstracts (FullAbs) on the same subset of the data. There are a total of 603 uncovered elements. 411 elements (68.16%) of these are stylistic content, 78 (12.94%) are due to tokenisation errors, 12 (1.99%) are due to errors made by MetaMap, and 29 (4.81%) are due to the use of generalisation in bottom-line summaries. We classify a total of 73 (12.11%) elements as missing content. As this number is much higher compared to the missing content from human-authored summaries, we further analyse the cause of this. The primary reason behind this is incomplete information in our corpus. 51 of the 73 uncovered elements are due to missing abstracts from the corpus or because of the PubMed abstracts missing text. As such, the text in some bottom-line summaries remain completely uncovered. Looking back at the histogram in Figure 6.3, this explains why the frequency in the left-most bin in the FullAbs is much higher (indicating close to zero coverage). When all the abstracts associated with a bottom-line summary are present in the corpus, generally they provide better coverage of non-medical content (often due to the greater volume of the source text).

QSpec. Following this, we study the uncovered elements from the QSpec summaries and random three sentence summaries on the same subset of the data. For the QSpec summaries, there are a total of 733 uncovered elements. 489 elements (66.71%) of these are stylistic content, 86 (11.73%) are due to tokenisation errors, 10 (1.36%) are due to errors made by MetaMap, and 34 (4.64%) are due to the use of generalisation in the bottom-line summaries. 113 (15.55%) elements were categorised as missing content. Since the automatic summaries are generated from only the abstracts in our corpus that have text, the majority of the content classified as missing are due to information missing from the corpus. Some relevant information is also lost during summarisation, causing the proportion of missing content to rise from 12.11% to 15.55%. However, proportion-wise, it can be seen that only a small proportion of relevant information is lost during the summarisation process, while at the same time, a lot of noise is removed. Thus, we can not reject the potential of using the automatic single-document summaries to investigate approaches for the generation of bottom-line summaries.

Random. For the random summaries, there are a total of 947 uncovered elements, which is significantly more than for QSpec summaries. As one would expect, a large number of the random summaries miss out on useful information. The category-wise distribution is as follows: 471 (49.74%) stylistic tokens, 106 (11.19%) cases of tokenisation and other forms of errors, 14 (1.48%) cases of MetaMap errors, 33 elements representing generalisation/specialisation, and as many as 323 (34.11%) elements categorised as missing content (as compared to 15.55% for QSpec). Figure 6.4 shows the distribution of the uncovered elements for the HS, FullAbs, QSpec and Random sets.

Figure 6.5 shows a random sample of uncovered elements from our data set for each category. Figure 6.6 (multi-page) provides a full example, showing the individual document summaries, the bottom-line summary and the uncovered elements. The full example is specifically chosen because it contains uncovered elements from each different category. The example shows 12 uncovered elements (stemmed). The tokens 'insulin.' and 'glucose-low' did not match because of tokenisation errors. The tokens 'avandia' and 'acto' (for *Actos*) represent specific commercial names that were added by the human authors. The summarised extracts, as one would expect, contain the generic names for these chemicals. Thus, these can be considered to be cases of generalisation. The tokens 'hb' and 'a1c' are not detected by MetaMap to represent the same concept and are therefore treated as separate tokens. The terms 'appear' and 'either' represent stylistic uncovered elements.

If the individual summaries are read, it can be seen that they represent the same conclusions as the bottom-line summary. The last sentence of the bottom-line summary states the absence of randomised trials comparing patient-oriented evidence. This information is implicitly present in the associated single-document summaries, but the automatic generation of such statements is beyond the scope of our research. This statement also results in 4 uncovered elements ('current', 'directli', 'patient-ori', 'outcom') that can be considered as missing information.

Section Conclusions

In this section, we presented a description of our investigations to estimate the extent to which the information in the bottom-line summaries are contained in source texts of various granularities. Our investigations suggest that a significant amount of the information in the bottom-line summaries comes from the source texts. The high coverage of human-authored summaries, which can be considered to be the best possible summaries of the source texts, indicate that a two-step summarisation approach shows promise for this task. In the previous chapter, we described the implementation of an extractive summariser that attempts to generate three sentence summaries that closely resemble the human-authored summaries in terms of content. Although our extractive summariser (QSpec) outperforms other state-of-the-art summarisers for this task,

Stylistic/linguistic tokens: `(', `)', also, include, low, alone, treat, may, use, 300, 600, rate, center, include, prevent, recommend, similar, adverse, effects, report, systematic, limit, inform, safety, box, among, statistically, significant, demonstrate, label, during, attain, likelihood, nation, education, program, added, provide, [,], number, increase, primarily, continue, limit, lower, settings, dosage, toxic, improve, clinical, course, add, all, produce, best, show, trade, loss, may, occur.

Tokenisation errors: sodium/sodium, effective., help., 3-month, mg/d, sub-analysis, therapy., study., low-density, lipid-modifying, statin/gemfibrozil, statin/ezetimibe, statin/bile acid sequestrant, 5-6, deficit/hyperactivity, short-term, attacks., first-line, 2.3, antidepressant - specifically; sertraline - are; prilocaine-lidocaine.

MetaMap errors: ssri *vs..* selective serotonin uptake inhibitors; triiodothyronine *vs..* t3; hbp *vs..* high blood pressure.

Generalisation/Specialisation: anticonvulsant for lamotrigine; atypical antipsychotic for quetiapine; ssri for fluoxetine; disease for bipolar depression; drug for various drugs; bile acid sequestrant for ezetimibe; fibrates for gemfibrozil; stimulant medication for methylphenidate; antidepressant for clomipramin, fluoxetine, and sertraline; PDE5 inhibitors for sildenafil, vardenafil, and tadalafil; tricyclic antidepressant for citalopram, escitalopram and venlafaxine; proton pump for omeprazole and esomeprazole; anti-fungals for fluconazol, griseofulvin; steroid for dexamethasone and acyclovir; thiazolidinediones for pioglitazone and rosiglitazone.

Missing information: lamotrigine adjunct; lithium; cure defined as 14 consecutive dry nights; inadequate evidence for oral *vs..* nasal form of desmopressin; injecting into tendon compartments more effective; may cause intolerable side effects; empiric therapy is appropriate; available data do not allow for adjusted risk assessment for patients with preexisting renal disease; chronic rhinosinusitis; likely to be more costly and cause side-effects; more tolerable.

Figure 6.5: Examples of *uncovered* elements from bottom-line summaries belonging to each category.

Full Example
Question: How beneficial are thiazolidinediones for diabetes mellitus?

Single-document Summaries:

1 [PMID: 11790216]: To systematically review available data from the literature regarding the efficacy of oral antidiabetic agents, both as monotherapy and in combination. Long-term vascular risk reduction has been demonstrated only with sulfonylureas and metformin. With few exceptions, the available oral antidiabetic agents are equally effective at lowering glucose concentrations.

2 [PMID: 11790217]: The scenarios also highlight some of the difficulties in choosing the optimal pharmacologic treatment regimen for individual patients. Physicians should also recognize that type 2 diabetes is a multisystem disorder that requires multidisciplinary care, including education and ongoing counseling for effective patient self-management of the disease. Finally, patient preferences are a vital component of informed decision making for pharmacologic treatment of diabetes.

3 [PMID: 10644273]: In controlled trials, there has been no evidence of rosiglitazone-induced hepatocellular injury. Clinical evaluation and assessment of liver function test results were done daily during hospitalization and periodically after discharge . We believe that patients receiving rosiglitazone should have liver enzyme levels monitored earlier and more frequently than initially recommended.

4 [PMID: 10644272]: Rosiglitazone maleate is the second approved oral hypoglycemic agent of the thiazolidinedione class. There have been no reports to date of rosiglitazone-associated elevations in the alanine aminotransferase level or hepatotoxicity. Discontinuation of rosiglitazone therapy and treatment with lactulose, vitamin K, fresh frozen plasma, ventilatory assistance, and intensive care unit support.

5 [PMID: 11232013]: After a 4-week placebo run-in period, 493 patients with type 2 diabetes were randomized to receive rosiglitazone [2 or 4 mg twice daily (bd)] or placebo for 26 weeks. Homeostasis model assessment estimates indicate that rosiglitazone (2 and 4 mg bd) reduced insulin resistance by 16.0% *and 24.6%, respectively, and improved ss-cell function over baseline by 49.5% and 60.0%, respectively. In the short-term, rosiglitazone is an insulin sensitizer that is effective and safe as monotherapy in patients with type 2 diabetes who are inadequately controlled by lifestyle interventions.*

6 [PMID: 10755495]: The complementary actions of the antidiabetic agents metformin hydrochloride and rosiglitazone maleate may maintain optimal glycemic control in patients with type 2 diabetes; therefore, their combined use may be indicated for patients whose diabetes is poorly controlled by metformin alone. Glycosylated hemoglobin levels, fasting plasma glucose levels, insulin sensitivity, and beta-cell function improved significantly with metformin-rosiglitazone therapy in a dose-dependent manner. Our data suggest that combination treatment with once-daily metformin-rosiglitazone improves glycemic control, insulin sensitivity, and beta-cell function more effectively than treatment with metformin alone.

7 [PMID: 10691158]: This study was designed to test the efficacy and safety of low-dose rosiglitazone, a potent, insulin-sensitizing thiazolidinedione, in combination with sulphonylurea in Type 2 diabetic patients. Both HDL-cholesterol and LDL-cholesterol increased and potentially beneficial decreases in non-esterified fatty acids and gamma glutamyl transpeptidase levels were seen in both rosiglitazone groups. Overall , the combination of rosiglitazone and a sulphonylurea was safe, well tolerated and effective in patients with Type 2 diabetes.

1 [PMID: 11423507]: To determine the efficacy and safety of rosiglitazone (RSG) when added to insulin in the treatment of type 2 diabetic patients who are inadequately controlled on insulin monotherapy. By intent-to-treat analysis, treatment with RSG 8 mg plus insulin resulted in a mean reduction from baseline in HbA(1c) of 1.2% *(P < 0.0001), despite a 12% mean reduction of insulin dosage. The addition of RSG to insulin treatment results in significant improvement in glycemic control and is generally well tolerated.*

2 [PMID: 11092281]: To evaluate the efficacy and safety of four doses of pioglitazone monotherapy in the treatment of patients with type 2 diabetes. Patients who had HbA1c > or = 7.0%, *fasting plasma glucose (FPG) > or = 140 mg/dl, and C-peptide > 1 ng/ml were randomized to receive placebo or 7.5, 15, 30, or 45 mg pioglitazone administered once a day for 26 weeks. Pioglitazone monotherapy significantly improves HbA1c and FPG while producing beneficial effects on serum lipids in patients with type 2 diabetes with no evidence of drug-induced hepatotoxicity.*

3 [PMID: 11192132:]: Their complimentary mechanisms of action suggest that a combination of pioglitazone hydrochloride and metformin may have clinically beneficial effects in the treatment of patients with type 2 diabetes. The pioglitazone + metformin group had significant mean percentage changes in levels of triglycerides (-18.2%) *and high-density lipoprotein cholesterol (+8.7%) compared with placebo + metformin (P < or = 0.05). In this study in patients with type 2 diabetes mellitus, pioglitazone + metformin significantly improved HbA1c and FPG levels, with positive effects on serum lipid levels and no evidence of drug-induced hepatotoxicity.*

4 [PMID: 11448655]: To evaluate the efficacy and tolerability of pioglitazone in combination with a sulfonylurea in the treatment of type 2 diabetes mellitus. Both pioglitazone + sulfonylurea groups had significant (P < 0.05) mean percent decreases in triglyceride levels (17%, *95% CI: 6% to 27% for 15 mg; 26%, 95% CI: 16% to 36% for 30 mg) and increases in high-density lipoprotein cholesterol levels (6%, 95% CI: 1% to 11% for 15 mg; 13%, CI: 8% to 18% for 30 mg) compared with placebo + sulfonylurea. In patients with type 2 diabetes, pioglitazone plus sulfonylurea significantly improves HbA1C and fasting plasma glucose levels with beneficial effects on serum triglyceride and HDL-cholesterol levels.*

Bottom-line Summary:

1 The thiazolidinediones pioglitazone (Actos) and rosiglitazone (Avandia) are effective at lowering fasting plasma glucose (FPG) and glycosylated hemoglobin (Hb A1c) in patients with type 2 diabetes when used either as monotherapy or in combination with sulfonylureas, metformin, or insulin. The glucose-lowering effects appear comparable with those of sulfonylureas and metformin alone. Currently, there are no randomized trials directly comparing patient-oriented outcomes of the thiazolidinediones with those of sulfonylureas and metformin.

Uncovered Elements (stemmed):

1 'insulin.', 'avandia', 'directli', 'glucose-low', 'either', 'current', 'patient-ori', 'outcom', 'a1c', 'acto', 'hb', 'appear'

Figure 6.6: Full example showing single-document summaries, the bottom-line summary, and the uncovered tokens.

it still falls short of the human-authored summaries in terms of coverage. However, analysis of the uncovered elements reveals that the difference between the coverage scores of the QSpec summary sentences and the full abstract is mostly due to linguistic/stylistic content. The QSpec summariser is capable of removing significant amounts of noise from the source texts with very little loss of useful information. Future improvements to the extractive summariser can perhaps further improve its performance. The performance of the summariser is dependent on the performance of a number of prior processes. They include, but are not limited to, the performance of MetaMap, the accuracy of the PIBOSO classification approach, and the accuracy of the question classification approach. We incorporate query information in our summariser only in the form of *query types* and the MMR algorithm. However, even for the same query type, the information needs of a question may be different. For example, for a *treatment and prevention question*, the information needs of a question beginning with the word *what* can be very different to the information needs of a question beginning with the word *when*. A detailed query analysis needs to be performed to better understand the information needs of queries and incorporate that information in the extractive summarisation approach. Furthermore, the size of our corpus is relatively small, with 456 questions. For several types of questions, there are only a few instances available in our corpus, and therefore, the statistics associated with those questions may not be completely reliable. Larger annotated data sets are likely to improve the performance of our extractive summarisation system. We also limit our summariser to selecting three sentences only, and it may not always be possible to incorporate all the useful information in three sentences. Finally, our summariser does not take into account the various *aspects* associated with a question. These *aspects* include duration of treatment, side effects, target population (*e.g.*, adolescents, adults), and so on. Considering the importance of these aspects, incorporating information regarding these is likely to generate better summaries. We discuss some of these later in this chapter. To summarise the findings of the experiments described in this section, we can say that the results support the hypothesis regarding a two-step approach, and that the QSpec extractive summaries show some promise for the generation of bottom-line summaries.

6.3 Study of Possible Approaches for Generation of Multi-document Summaries

In this section, we discuss possible approaches for synthesising the key information from single-document summaries to generate bottom-line summaries. The generation of bottom-line summaries requires the synthesis of information from various sentences presenting outcomes that are associated with the same clinical query. The process of transformation of summary sentences to produce bottom-line summaries can be considered to be an abstractive multi-document summarisation task. This is because the bottom-line summaries simply make recommendations that

are derived from the information present in the individual documents, and the actual texts from the individual documents are not required in the final summaries. In this section, we analyse the applicability of two broad categories of multi-document summarisation approaches for our task. The two categories of approaches are:

1. Redundancy-reliant approaches — perhaps the most common approaches to multi-document summarisation, whether extractive or abstractive, are those that exploit redundant information across text segments in distinct documents to identify key information which should be included in summaries. Approaches reliant on the presence of redundant information have been successfully applied to domains where text from distinct documents are similar. For example, in the news domain, where distinct articles on the same topic/event present the same information in different ways, such redundancy-based approaches have been particularly popular. We study the *sentence fusion* approach [Barzilay and McKeown, 2005] as a representative, state-of-the-art, redundancy-reliant approach, and compare the data used for this task with the data in our corpus to assess the applicability of such approaches for performing information synthesis using the single-document summaries we have.

2. Polarity Classification [Niu et al., 2006] — this is an abstractive summarisation approach that attempts to identify whether a sentence/document represents positive or non-positive information in a given context (*e.g.*, via a query). This approach is representative of opinion summarisation and sentiment summarisation approaches, and has the potential of being applied in a multi-document scenario, where the polarities of multiple text segments can be combined to generate a final polarised recommendation. The key challenge in this approach is the automatic identification/classification of polarities of individual sentences, which can then be utilised for the generation of a bottom-line answer.

We carried out preliminary analyses on both these classes of approaches to determine their suitability to our intended task. We provide details of our analyses in the following subsections.

6.3.1 Redundancy-reliant Approaches: Summarisation via Sentence Fusion

The idea of sentence fusion for summarisation was first proposed by Barzilay and McKeown [2005], and this approach was shown to be effective for summarising news articles. The intent of sentence fusion is to convert multiple sentences that have high redundancy (overlap) of information into a single summary sentence with reduced redundancy. All the sentences are tightly clustered around a single news event, and therefore, can be identified using existing clustering algorithms. The common information presented by similar sentences in different documents are first identified to realise a general proposition in the summary sentence, by a

process called *sentence alignment*. Following that, the proposition is *augmented* by adding relevant information from various sentences and *pruned* for compression.

The primary reason behind the successful application of the sentence fusion algorithm is the amount of redundant information present among the text in the news domain. Sentences from different news articles on the same topic almost invariably contain some common information. Figure 6.7 gives two examples of similar sets of sentences taken from different news articles (the first example has been taken from Barzilay and McKeown [2005]; for the second example, the sources are mentioned). It can be clearly seen that at least two of the three sentences in the first example contain some key common information about the incident, such as: 'Palestinian/Palestinian militant', 'fire', 'antitank missile/machine gun and antitank missile', and 'bulldozer'. With the sentence fusion algorithm, these key elements are used for aligning the sentences, and the summary sentence may be augmented with additional information (*e.g.*, 'Israeli forces', 'army base', or 'area'). For both examples, the different common elements in each sentence, and the final fused sentence, are specified using different colours in the figure.

The sentence fusion algorithm works successfully on text from the news domain because of this high degree of similarity between sentences from different articles, and the existence of key common (redundant) information. The high degree of similarity among the sentences presenting similar information is also important for this approach because it enables the successful clustering of the sentences. Numerous other multi-document summarisation techniques, particularly redundancy-reliant ones, utilise the same property (*i.e.*, highly similar contents in sentences from multiple documents) for summarisation.

In the domain of evidence-based medicine, multi-document summarisation requires information synthesis from sentences located in distinct documents that contain relevant information associated with a clinical query. For example, the documents may discuss the effects of different drug therapies for treating a specific disorder; or the different etiologies for a condition. Our analysis of the QSpec summary sentences revealed that the summary sentences associated with the same bottom-line answer do not exhibit the same degrees of inter-sentence similarities as those from news texts. One of the key reasons behind this phenomenon is the fact that the summary sentences are not associated with the same event, unlike news sentences. As such, these summary sentences generally present information in different ways, and the way in which the information is expressed often depends on the intents and topics of the source articles. Therefore, alignment of these sentences for performing sentence fusion is more challenging, and *may* also not be conceptually meaningful. Figure 6.8 shows three sentences that are taken from separate documents as answers to the question: 'Is antibiotic prophylaxis effective for recurrent acute otitis media?'. From the figure, it can be seen that, unlike news sentences, these sentences are quite distinct from each other, although they all convey the limited effectiveness of antibiotics for acute otitis media.

Example 1:

1. IDF Spokeswoman did not confirm this, but said the Palestinians fired an antitank missile at a bulldozer.
2. The clash erupted when Palestinian militants fired machine guns and antitank missiles at a bulldozer that was building an embankment in the area to better protect Israeli forces.
3. The army expressed “regret at the loss of innocent lives" but a senior commander said troops had shot in self-defense after being fired at while using bulldozers to build a new embankment at an army base in the area.

   Fused Sentence: : Palestinians fired an antitank missile at a bulldozer.

Example 2:

1. Israeli soldiers shot dead a 21-year-old Palestinian woman near the West Bank city of Hebron on Wednesday and wounded another local youth. (Reuters)
2. A 21-year-old Palestinian woman has died after being shot in the face by Israeli soldiers in the West Bank. (Al-Jazeera)
3. Israeli soldiers shot and killed a 21-year-old Palestinian woman near the West Bank city of Hebron on Wednesday. (Yahoo! News)
4. A 21-year-old Palestinian woman died after being hit in the face by Israeli gunfire in the West Bank on Wednesday, medics said, with witnesses saying she was shot by soldiers. (AFP)
5. An Israeli soldier driving a civilian car shot and killed a 21-year-old Palestinian woman Wednesday near a refugee camp outside the West Bank city of Hebron, according to Palestinian witnesses and medical sources. (Los Angeles Times)

   Fused Sentence: : Israeli soldiers shot a 21-year-old Palestinian woman near the West Bank.

Figure 6.7: Similar sentences from news articles, used for summarisation using sentence fusion. Example 1 taken from Barzilay and McKeown [2005]. The sources for Example 2 are shown in the figure.

1. Antibiotics appear to have beneficial but limited effect on recurrent otitis media and short-term resolution of OME.
2. Because of the potential of excessive antibiotic use to promote the acquisition of resistant pneumococci and the lack of effectiveness in this trial, routine use of amoxicillin prophylaxis should be discouraged.
3. Selective use of tympanocentesis if the patient does not respond to empiric therapy can help confirm the diagnosis and guide effective therapy.

Figure 6.8: Summary sentences from distinct medical documents expressing the limited effectiveness of antibiotics for a specific treatment.

Generating a bottom-line, multi-document summary by fusing these sentences together would be an extremely challenging task, especially for existing text-to-text summarisation algorithms (*e.g.*, sentence fusion) that rely on the availability of redundant information.

To investigate our preliminary findings about the possible inapplicability of the sentence fusion algorithm to text in the evidence-based medicine domain, we now present a brief quantitative analysis. For this analysis, we collected the data used by Barzilay and McKeown [2005] for their sentence fusion task. In their work, clusters of *potentially fusable* sentences are first identified and ranked, and the top ranked clusters are used for sentence fusion. We use a sample of 110 sets of *potentially fusable* groups of sentences (prepared by Barzilay and McKeown [2005]) for our experiments. Unlike the authors, however, we do not perform any ranking of these groups of sentences, and instead, choose all the 110 sets. We do this intentionally, to ensure that we were not adding bias to our experiments by only using groups of highly similar sentences. Also, in our domain, if sentence fusion or any other redundancy-reliant approach is to be applied, this ranking step has to be skipped. This is because successful application of the approach in the evidence-based medicine domain will require the sentences extracted as the query-focused, single-document summaries to be fusable, irrespective of how the sentences rank in a similarity measure.

In order to determine the applicability of redundancy-reliant approaches for evidence-based medicine summarisation, we compare the *similarities* among *potentially fusable* news sentences to the *similarities* among the single-document summary sentences (QSpec summaries) from which multi-document summaries are to be generated. We use two similarity measures that we have applied in the previous chapter: *cosine similarity* and *jaccard similarity*. We used 110 sets of news sentences and 600 sets of single-document summary sentences (each set associated with

a single bottom-line summary). To compute the similarities among sentences in each set, we apply the following approach:

1. Given a set of sentences, compute the top 25% *tf.idf* terms in that set of sentences. These top 25% *tf.idf* terms are used to represent the *centroid* of the set of sentences. *tf* here represents the frequency of a term in a cluster, and *df* represents the number of clusters containing that term. *idf*, therefore, is the inverse of this value.
2. For each sentence in the set, compute its *jaccard* and *cosine* similarities with the centroid.
3. Compute the average similarity across all the sentences in the set.

Using this approach, we compute the average *cosine* and *jaccard* similarities for all the sets of sentences in both domains. For the 110 sets of news sentences, the average *cosine* similarity across all the sets is 0.36, and the average *jaccard* similarity across all sets is 0.21. In comparison, the average similarities across the 600 clusters of evidence-based medicine sentences are much lower: *cosine* = 0.24 (0.12 units lower than news); *jaccard* = 0.14 (0.07 units lower than news). To give the reader an idea of how much lexical similarity these values actually indicate: for the second example in Figure 6.7, the average cosine similarity is 0.68 and the average jaccard similarity is 0.47; for the example in Figure 6.8, the average cosine similarity is 0.53 (0.15 units lower than the news example) and the average jaccard similarity is 0.31 (0.16 units lower than the news example). In both cases, the evidence-based medicine text shows about two-thirds of the similarity of the news text. Figure 6.9 shows the distributions of the average similarities among sentences from these two domains. The distributions illustrate the higher similarities among sentences in the news domain, compared to the evidence-based medicine domain – thus quantitatively illustrating why redundancy-reliant summarisation approaches are likely to be less effective in this domain. The distributions also show that, for the news domain, some clusters of sentences tend to have quite high average similarities.

We perform another set of experiments to check if there is a significant difference in the average similarities among sentences in the news domain compared to the medical domain. We are interested in the similarities among all sentences associated with a specific topic. Therefore, for the news domain, we measure the average similarities for all sentences belonging to all articles associated with a news event. For the sentence fusion data set, there are 16 such events, with a total of 106 articles associated with them (average number of articles per event: 6.625). For the evidence-based medicine domain, we measure the average similarities for all the sentences belonging to abstracts associated with a bottom-line summary. In an attempt to ensure that the two data sets are comparable, we only include bottom-line summaries that have at least three abstracts associated with them. In total, there are 136 such bottom-line summaries with a total

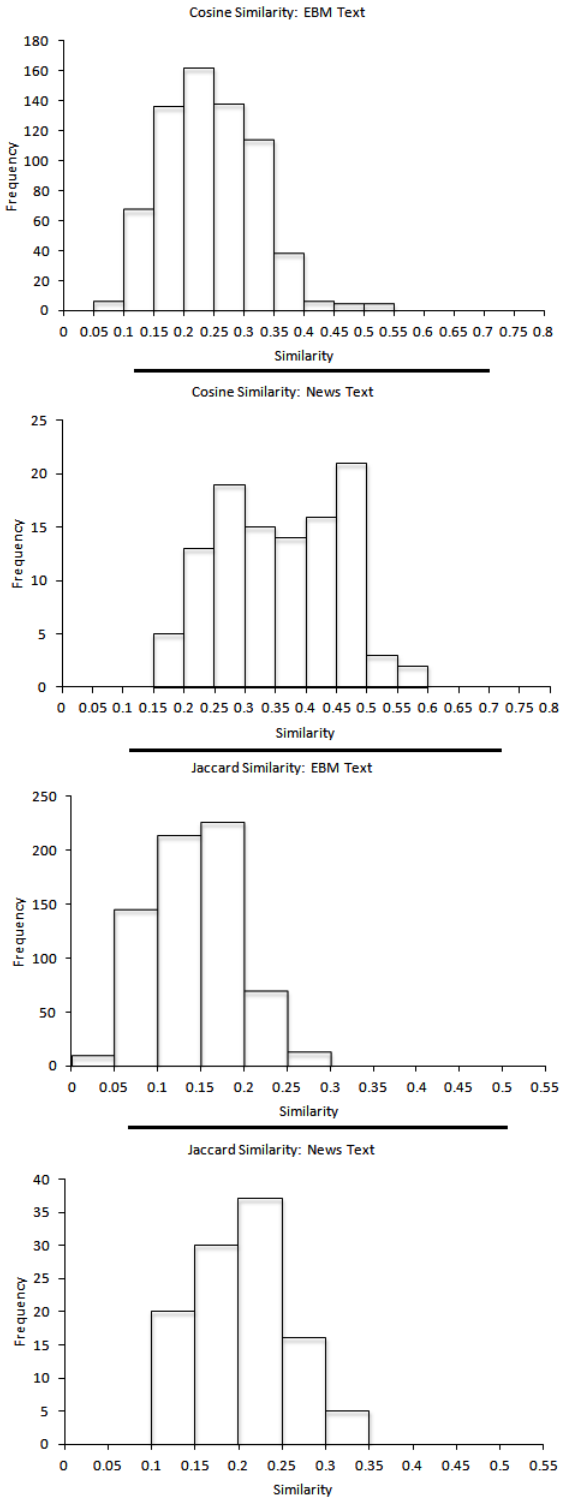


Figure 6.9: Comparison of the cosine and jaccard similarity distributions for news text and evidence-based medicine text.

of 760 associated abstracts (average number of abstracts per summary: 5.588). For the news data set, the average sentence similarities for all the sentences associated with a specific event are: cosine similarity – 0.137; jaccard similarity – 0.058. For the evidence-based medicine data set, the average sentence similarities for all sentences associated with a single clinical query are much lower: cosine similarity – 0.073, jaccard similarity – 0.030. To compute these similarity measures, each sentence is compared with the centroid for the associated set of sentences.

We also compute the similarity for each sentence compared to all the terms in the set (as opposed to the centroid of that set). Note that, in this case, the jaccard similarity is not a good measure because longer sentences will always have more overlap, and therefore higher similarity scores. For this, the cosine similarity for news text is 0.361, and for evidence-based medicine text is 0.293. It can be seen that, in all the results so far, the similarity values for news text are higher than those for the evidence-based medicine text. We perform unpaired T tests to check if the differences in the values for average similarities are statistically significant. In the tests, our null hypothesis is that the average similarities are the same for both types of text. For all the 3 pairs of similarity values mentioned above, the one tailed p-value is less than 0.0001, meaning that the mean similarity scores are statistically significantly different, with those for the medical text being lower.

Further experimentation is required to investigate the differences among the similarities of text in the two domains. We leave this for future work and focus on the analysis of other approaches for the generation of bottom-line evidence-based medicine summaries. For now, based on the relatively low similarity values among the single-document summary sentences compared to clusters of fusable news sentences, the prospects of the sentence fusion (or other redundancy-reliant approaches) does not look encouraging for further exploration. Therefore, we move our attention to another class of approaches which appears to be more favourable for this task.

6.3.2 Polarity Detection-based Approaches: Summarisation via Sentence Polarity Classification

The bottom-line summaries in our corpus present final recommendations in response to queries. For example, a bottom-line summary may or may not recommend an intervention in response to a disorder. Thus, the bottom-line summaries can be considered to be polarised — when an intervention is recommended, the polarity is positive, and when it is not recommended, the polarity is non-positive. The bottom-line summaries, as explained earlier, are generated by synthesising information from individual documents. Therefore, it is expected that the polarities of the individual documents, or their summaries, agree with the polarities of the associated bottom-line summaries. In this subsection, we study the use of sentence level polarities to determine the final polarities of the bottom-line summaries.

Data and Annotation

For preliminary annotation and analysis, we use 33 manually identified questions from our corpus. All these questions are treatment questions, and the bottom-line summaries mention one or more interventions, some of which are recommended while the others are not. Our first step in preparing for this task is to annotate the bottom-line summaries. From each bottom-line answer associated with these 33 questions, we manually identify the interventions. Following that, we annotate the polarity of the answer relative to the interventions mentioned. We use two categories for the annotation:

i **Positive:** We annotate the summary to be positive relative to an intervention when the summary clearly recommends the use of an intervention or states its effectiveness.

ii **Non-positive:** We annotate a summary to be non-positive relative to an intervention when the summary (a) clearly states that the intervention is not recommended, (b) states that the intervention is harmful, (c) states that there is insufficient evidence to recommend the specific intervention, or (d) mentions an intervention but provides no useful information about it.

These two categories cover all the cases in our data. Furthermore, considering the fact that the purpose of the bottom-line summaries is to recommend or not recommend an intervention, we find the use of two categories to be sufficient. In total, 111 interventions are identified and annotated into these two categories.

Figure 6.10 presents two questions, the associated bottom-line summaries, and our contextual polarity annotations. In the bottom-line summary associated with the first question, the three *beta-blockers* are recommended, and they are annotated as such. The second example gives an example of a question and an associated bottom-line summary that does **not** recommend the use of *hormonal therapies* and *progesterone*, and they are annotated as such. All the answers to the 33 questions are annotated by two annotators, including the author of this thesis. In almost all the cases, there is no disagreement at all between the annotators; the few disagreements are resolved via discussion.

Next, we collect all the abstracts associated with each bottom-line summary and question pair. The full abstracts may, and generally do, contain noisy information that are not related to the query. Also, our coverage analysis (Section 6.2) showed that the summaries provided by the QSpec summariser are rich in content. So, we consider the QSpec summary sentences to be *key* sentences from the abstracts, and perform polarity annotation on the key sentences. Similar to our bottom-line summary annotation process, for a sentence, we first identify the intervention(s)

Question: What is the most effective beta-blocker for heart failure?

Bottom-line answer: Three beta-blockers-carvedilol, metoprolol, and bisoprolol-reduce mortality in chronic heart failure caused by left ventricular systolic dysfunction, when used in addition to diuretics and angiotensin converting enzyme (ACE) inhibitors.

Contextual Polarities: carvedilol – recommended; metoprolol – recommended; bisoprolol – recommended.

Question: What medications are effective for treating symptoms of premenstrual syndrome PMS?

Bottom-line answer: Hormonal therapies (oral contraceptives, gonadotropin-releasing hormone agonists, danazol, estrogen) lack convincing evidence of efficacy and cause many side effects; progesterone is no more beneficial than placebo.

Contextual Polarities: hormonal therapies – not-recommended; progesterone – not-recommended.

Figure 6.10: Sample bottom-line summaries and examples of polarity annotations.

mentioned, and then categorise their polarities. Due to this annotation approach, we come across sentences where two different interventions are mentioned and the polarities associated with them are opposite. Consider the following sentence, for example:

> *The present study demonstrated that the combination of cimetidine with levamisole is more effective than cimetidine alone and is a highly effective therapy for the treatment of recalcitrant warts.*

For this sentence, the combination of two drugs is recommended over monotherapy with *cimetidine*. Therefore the polarities are: *cimetidine with levamisole* – recommended; *cimetidine alone* – not recommended. At the same time, in a number of cases, although a sentence is polarised, it does not mention an intervention. Such sentences are annotated without adding any intervention name to the context. The following is an example:

> *One injection cured 64% [corrected] of patients with primary trigger finger with no side effect and is the recommended nonsurgical treatment.*

In this sentence, *injection* refers to *steroid and lidocaine*, which are mentioned in the previous summary sentence. However, at this point, we are not interested in complicating the process by introducing inter-sentence relationships. So we leave this sentence annotation intervention-context free. In this manner, we annotate a total of 589 sentences from the QSpec summaries associated with the 33 questions. If a sentence contains more than one intervention, we add an annotation for each intervention.

A subset of the QSpec sentences, 124 in total, are annotated by a second annotator, and these annotations are used to measure agreement among the annotators. We use the Cohen's Kappa [Carletta, 1996] measure (shown below) to compute inter-annotator agreement.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{6.2}$$

where $Pr(a)$ is the relative observed agreement among annotators, and $Pr(e)$ is the hypothetical probability of chance agreement. We obtained $\kappa = 0.85$, which can be regarded as almost perfect agreement [Landis and Koch, 1977].

Analysis

Following the annotation process, we compare the annotations of the single-document summary sentences with the bottom-line summary annotations. Given that a summary sentence has

been annotated to be of positive polarity with an intervention in context, we first check if the intervention (or a generalisation of it) is also mentioned in the bottom-line summary. If yes, we check the polarity of the bottom-line summary. In this manner, we collect a total of 177 summary sentence – bottom-line summary pairs. Among these, in 169 (95.5%) cases, the annotations are of the same polarity. In the rest of the 8 cases, the QSpec summary sentence recommends a drug, but the bottom-line summary does not.

We manually examine the 8 cases where there are disagreements between the single-document summary sentence annotations and the bottom-line summary annotations. In all the cases, this is either because individual documents present contrasting results, that is, the positive findings of one study were negated by evidence from other studies; or because a summary sentence presents some positive outcomes, but side effects and other issues are mentioned by other summary sentences, eventually leading to an overall negative polarity. Consider the following example sentence, taken from a QSpec summary:

> *By the end of the trial more than 75% of patients who were still taking danazol were essentially free of breast pain, lethargy, anxiety and increased appetite, but results for other common symptoms were no better than with placebo.*

The sentence was generated as part of a summary in response to the question: *What medications are effective for treating symptoms of premenstrual syndrome PMS?* This sentence presents some positive outcomes, and therefore it is annotated to be of positive polarity (with *danazol* as the context intervention). However, another summary sentence associated with the same topic, but from a different document, clearly states that the drug in question is not effective. The sentence is as follows:

> *Luteal phase-only danazol is not effective for the treatment of the general symptoms of premenstrual syndrome but appears highly effective for the relief of premenstrual mastalgia.*

Thus, the bottom-line summary does not recommend the use of *danazol*, causing the disagreement in the two sets of annotations.

If automatic sentence level polarity classification techniques are to be used for generating bottom-line summaries in a two-step summarisation process, the first step (QSpec summaries) also needs to have very good recall. For the data set used in this analysis, the bottom-line summaries mention a total of 111 interventions (of both polarities). Of them, the QSpec summary sentences contain 100, giving a recall of 90.1%. We examine the causes for unrecalled drug names and find that, of

the 10 drug names not recalled, 4 are due to missing abstracts from the corpus, and 2 drug names are not mentioned in any of the referenced abstracts. Thus, the actual recall is 95.2%.

Considering the high recall of intervention names in the summary sentences, and the high agreement among the QSpec summary sentence polarities and bottom-line summary sentence polarities, it appears that automatic polarity classification techniques have the potential to be applied for the task of bottom-line summary generation in a two-step summarisation process. Understandably, preliminary automatic summarisation work in this domain has already explored the idea [Niu et al., 2006]. We now discuss our approach to automatic sentence level polarity classification.

Automatic Polarity Classification

We model the problem of automatic sentence level polarity classification as a supervised classification problem. We annotate a total of 2,362 key sentences (QSpec summaries) from the corpus (1,736 non-positive and 626 positive instances). For each sentence, we first identify the intervention(s) mentioned in it, and the polarity of the sentence relative to each intervention. Given a sentence and an intervention mentioned in the sentence (if any), our intent is to train a machine learning algorithm to automatically identify the polarity of the sentence relative to the intervention. The same sentence may appear as multiple instances, each time with a different context intervention, and a polarity class associated with the intervention. Our input, therefore, to the classifier are: a sentence and a context intervention (if present), and the expected output is the polarity of the sentence relative to the context intervention. If the sentence contains no interventions, the output is simply the context-free polarity of the sentence.

We automatically extract a number of features from each sentence. Some of these features are context-sensitive, meaning that the same sentence may have different feature values associated with it based on the intervention in context. Because there has been some past work in sentence level polarity classification in this domain (as discussed in Chapter 2), we build on the features proposed by existing research on sentence level polarity classification and add some previously unexplored features: context-specific and context-independent. The following is a description of the features.

1. Word n-grams

Our first feature set is word n-grams (n = 1 and 2) from the sentences. Cues about the polarities of sentences are primarily provided by the lexical information in the sentences (*e.g.*, words and phrases). As such, word n-grams is the most natural choice for a feature set. We perform some preprocessing of each sentence. We lowercase the words, remove stop words and stem the words using the Porter stemmer [Porter, 1980]. For each sentence that has an annotated context, we

replace the context word(s) using the keyword '_CONTEXT_'. This ensures that our classifiers are not influenced by the names of the interventions. Furthermore, we replace the disorder terms in the sentences using the keyword '_DISORDER_'. To identify the various disorder terms in the sentences, we use MetaMap. We use the same semantic types as Uzuner et al. [2009][5] as the disorders.

2. Change Phrases

We use the Change Phrases features proposed by Niu et al. [2005]. The intuition behind this feature set is that the polarity of an outcome is often determined by how a change happens: if a *bad* thing (*e.g.*, mortality) was *reduced*, then it is a positive outcome; if a *bad* thing was *increased*, then the outcome is negative. This feature set attempts to capture cases when a *good/bad* thing is *increased/decreased*. To construct these features, we first collect the four groups of *good*, *bad*, *more*, and *less* words used by Niu et al. [2005]. We augment the list by adding extra words to the list which we expect to be useful. In total, we add 37 *good*, 17 *bad*, 20 *more*, and 23 *less* words. The full list of words is presented in Appendix D. Our intent with this feature set is to utilise the co-occurrence of more/less words and good/bad words to detect the *change in polarity*. This feature set has four features: MORE-GOOD, MORE-BAD, LESS-GOOD, and LESS-BAD. As an example, the following sentence exhibits the LESS-BAD feature, indicating a positive polarity.

> *Statistically and clinically significant improvement, including a statistically significant* ***reduction*** *in* ***mortality****, has been noted in patients receiving therapy with either bisoprolol, carvedilol, or metoprolol.*

To extract the first feature, we follow the approach applied by Niu et al. [2005]: a window of four words on each side of a MORE-word in a sentence is observed. If a GOOD-word occurs in this window, then the feature MORE-GOOD is activated (its value is set to 1). The other three features are activated in a similar way. The features are represented using a binary vector with 1 indicating the presence of a feature and 0 indicating absence.

3. UMLS Semantic Types

We use all the UMLS semantic types (identified using MetaMap) present in a sentence as features. Intuitively, the occurrences of semantic types, such as *disease or syndrome* and *neoplastic process*, may be different in different polarity of outcomes. There is a possibility that the polarity of a sentence has some relation to the semantic types contained by it. Overall, UMLS provides 133 semantic types, and we represent this feature set using a binary vector of size 133 – with 1 indicating the presence and 0 indicating the absence of a semantic type.

[5]Semantic types in this category include: pathological function, disease or syndrome, mental or behavioral dysfunction, cell or molecular dysfunction, virus, neoplastic process, anatomic abnormality, acquired abnormality, congenital abnormality and injury or poisoning.

4. Negations

Negations play an important role in determining the polarity of the outcomes presented in medical sentences. For example, consider the following statements:

1. Drug X *significantly improves* quality of life.

2. Drug X *does not significantly improve* quality of life.

It can be clearly seen that the polarities of the two statements are completely opposite due to the presence of the negation term *'not'*. We attempt to detect the negation terms in the sentences and also automatically mark the negated terms in each sentence. For this, we apply three different negation detection strategies.

In our first strategy, we detect the negations using the same approach as Niu et al. [2005]. In their simplistic approach, the authors use the *'no'* keyword as a negation word and use that for detecting negated concepts. To extract the features, all the sentences in the data set are first parsed by the Apple Pie parser[6] to get phrase information. Then, in a sentence containing the word *no*, the noun phrase containing *no* is extracted. Every word in this noun phrase except *no* itself is attached a 'NO' tag. We use a similar approach, but instead of the Apple Pie parser, we use the GENIA Dependency Parser (GDep)[7] [Sagae and Tsujii, 2007], since it has been shown to give better performance with medical text.

For the second strategy, we use the BioScope corpus[8] [Vincze et al., 2008] to identify negations. The BioScope corpus contains various sentence level annotations including negations and their scopes. We do not take into account the scopes of negations but identify the negation terms from it. Then we apply the same strategy as before, using the GDep parser again.

For the third strategy, we use the NegEx system [Chapman et al., 2001] to identify negations. The NegEx system uses a list of negation terms and applies regular expression based matching to identify the negation terms in text. Using NegEx, we use the same strategy as above to identify and mark negated terms using the GDep parser.

5. PIBOSO Category of Sentences

In Chapter 5, we presented the use of sentence classification in extractive summarisation. Our analysis of the QSpec summary sentences suggests that the class of a sentence may be related to the presence of polarity in the sentence. For example, a sentence classified as *Outcome* is more

[6] http://nlp.cs.nyu.edu/app/. Accessed on 26th May, 2014.
[7] http://people.ict.usc.edu/~sagae/parser/gdep/. Accessed on 26th May, 2014.
[8] http://www.inf.u-szeged.hu/rgai/bioscope. Accessed on 26th May, 2014.

likely to contain a polarised statement than a sentence classified as *Background*. Therefore, we use the PIBOSO classifications of the sentences as a feature.

6. Synset Expansion
Certain terms play an important role in determining the polarity of a sentence, irrespective of context (*e.g.*, some of the *good* and *bad* words used in the *change phrases* feature). Certain adjectives, and sometimes nouns and verbs, or their synonyms, are almost invariably associated with positive or non-positive polarities. Thus, for each adjective, noun or verb in a sentence, we use WordNet[9] to identify the synonyms of that term and add the synonymous terms, attached with the 'SYN' tag, as features.

7. Context Window
This is the first of our context-sensitive features. We noticed that, in a sentence, the words in the vicinity of the context intervention may provide useful information regarding the polarity of the sentence relative to that drug. Thus, we collect the terms lying inside three word boundaries before and after the context drug term(s). This feature is useful when there are direct comparisons between two interventions. Consider the following example:

> *All three niacin-containing treatments were more effective than lovastatin monotherapy in reducing lipoprotein.*

There are two interventions mentioned: *niacin-containing treatments* and *lovastatin monotherapy*, the former having a positive contextual polarity and the latter having a non-positive one. The three-word window before *niacin-containing treatments* contains the terms *'all'*, and *'three'*, and the three-word window after the mention of the intervention contains the terms *'were'*, *'more'*, and *'effective'*. For *lovastatin monotherapy*, the three-word window before contains the terms *'more'*, *'effective'*, *'than'*, and the window after contains the terms *'in'*, *'reducing'*, *'lipoprotein'*. From the example, it can be seen that the terms appearing before an intervention and those appearing after it may play a role in determining the polarity of the sentence relative to a given context. We tag the words appearing before an intervention with the 'BEFORE' tag and those appearing after with the 'AFTER' tag and use these as features.

8. Dependency Chains
In some cases, the terms that influence the polarity of a sentence associated with an intervention do not lie close to the intervention itself, but are connected to it via dependency relationships. In order to capture terms that may be associated with an intervention via dependency relationships, we used the sentence parses produced by the GDep parser. For each intervention appearing in a sentence, we identify all the terms that are connected to it via specific dependency chains. We

[9] http://wordnet.princeton.edu/. Accessed on 26th May, 2014.

Sentence: *MRI has proved to be an excellent tool in diagnosing injuries of the cruciate ligaments and menisci.*

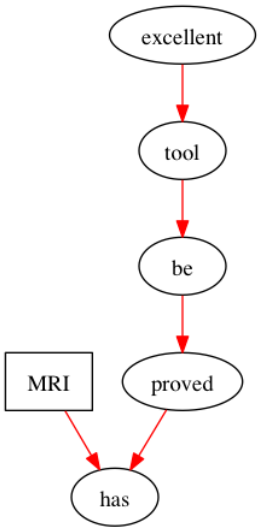


Figure 6.11: Dependency chain connecting an intervention (MRI) with a polarity influencing word (*excellent*).

use a rule to determine the chain, which is as follows:

1. Start from the intervention and move up the dependency tree till the first VERB item the intervention is dependent on, or the ROOT element.

2. Find all items dependent on the VERB item (if present) or the ROOT element.

All the terms connected to the context term via this relationship are collected, tagged using the 'DEP' keyword and used as features. Figure 6.11 gives us an example when this feature is useful. In the sentence shown, MRI is the context intervention. The *context window* feature does not capture the key polarity word in this sentence (*excellent*). However, the above algorithm is able to capture the fact that the term 'excellent' is associated with the intervention, and utilise that for the polarity classification. This feature is particularly useful for very long sentences, where the distance between key polarity-determining terms and the associated intervention may be large.

9. Other Features

We use a number of other simple numeric or binary features, which are listed below:

- context intervention position: This is a numeric feature that specifies the position of the intervention in the sentence (numeric);

- summary sentence position (numeric);
- presence of modals in the sentence (binary);
- presence of comparatives in the sentence (binary); and
- presence of superlatives in the sentence (binary).

Evaluation and Results

In our experiments, we use approximately 85% of our annotated data (2,008 sentences) for training and the rest (354 sentences) for evaluation. We performed preliminary 10-fold cross validation experiments on the training set using a range of classifiers and found Support Vector Machines (SVM) to give the best results, in agreement with existing research in this area. We use the SVM implementation provided by the Weka machine learning tool.[10]

Table 6.4 presents the results of our polarity classification approach. The overall accuracy obtained using various feature set combinations is shown, along with the 95% confidence intervals[11], and the F-scores for the positive and non-positive classes. The first set of features shown on the table represents the features used by Niu et al. [2006]; we consider the scores achieved by this system as the baseline scores. The second row presents the results obtained using all context-free features. It can be seen from the table that the two context-free feature sets, expanded synsets and PIBOSO categories, improve classification accuracy from 76% to 78.5%. This shows the importance of these context-free features. When all features are added, classification accuracies reach over 80%. The table shows the results for the three variants of negation detection, with the approach using the BioScope corpus giving the best accuracy. However, the differences between the different negation detection systems are minimal. All three variants give statistically significant increases in accuracy compared to the baseline.

Feature sets	Accuracy (%)	95% CI	Pos. F-score	Non-pos. F-score
1,2,3, and 4 (Niu)	76.0	71.2 – 80.4	0.58	0.83
Context-free (1-6)	78.5	73.8 – 82.8	0.64	0.85
All (Niu)	83.9	79.7 – 87.6	0.71	0.89
All (Bioscope)	84.7	80.5 – 88.9	0.74	0.89
All (NegEx)	84.5	80.2 – 88.1	0.73	0.89

Table 6.4: Polarity classification accuracy scores, 95% confidence intervals, and class-specific F-scores for various combinations of feature sets.

[10] http://www.cs.waikato.ac.nz/ml/weka/. Accessed on 26th May, 2014.

[11] Computed using the binom.test function of the R statistical package (http://www.r-project.org/. Accessed on 26th May, 2014.)

In terms of the F-scores, the table shows that the non-positive class F-scores are much higher than the positive class F-scores. The highest F-score obtained for the positive class is 0.74, and that for the non-positive class is 0.89. This is expected due to the fact that the number of training examples for the latter class is almost three times that of the positive class.

To explore the effect of the size of the training data on the classification accuracy and F-scores, and to predict how classification accuracies are likely to change if more training data is available, we perform several classification experiments keeping the test set constant, but varying the size of the training set. We use different sized subsets of the training set: starting from 10% of its original size and increasing the size by 5% each time. To choose the training data for each experiment, we perform random sampling with no replacement. Figure 6.12 illustrates the effect of the size of the training data on classification accuracies.

As expected, classification accuracies and F-scores increase as the number of training instances increases. The increase in the F-scores for the positive class is much higher than the increase for the non-positive class F-scores. This verifies that the positive class, particularly, suffers from the lack of available training data. Having more annotated data would invariably improve on the best F-score obtained in our experiments. Furthermore, none of the three curves flatten completely towards the right side of the figure, although the non-positive class F-scores and the overall accuracies seem to flatten more compared to the positive class F-scores. The increasing values for all three curves indicate that, if more training data are available, better results can be obtained for both classes. This is particularly true for the positive class, which is also perhaps the more important class considering our goal of generating bottom-line recommendations for clinical queries. The highest accuracy obtained by our system is 84.7%, which is significantly better than the baseline system for this domain.

To conclude this investigation, we perform a manual evaluation on a subset of the data to validate the suitability of the polarity classification approach for the generation of bottom-line summaries. For this evaluation, we use the same set of sentences that was used for the preliminary analysis — key sentences associated with the 33 manually identified questions. As mentioned earlier, a total of 111 interventions were identified from the human-authored bottom-line summaries and annotated. Our intent in this evaluation is to determine the extent to which bottom-line recommendations relative to specific interventions can be derived from the automatically classified sentences. We do not attempt to combine information from multiple sentences associated with the same context, and instead rely on the following simple rules:

1. If the polarity of a sentence, relative to a specified context intervention, is classified as positive, the intervention is considered to be a recommendation.

2. If the polarity of a sentence, relative to a specified context intervention, is classified as

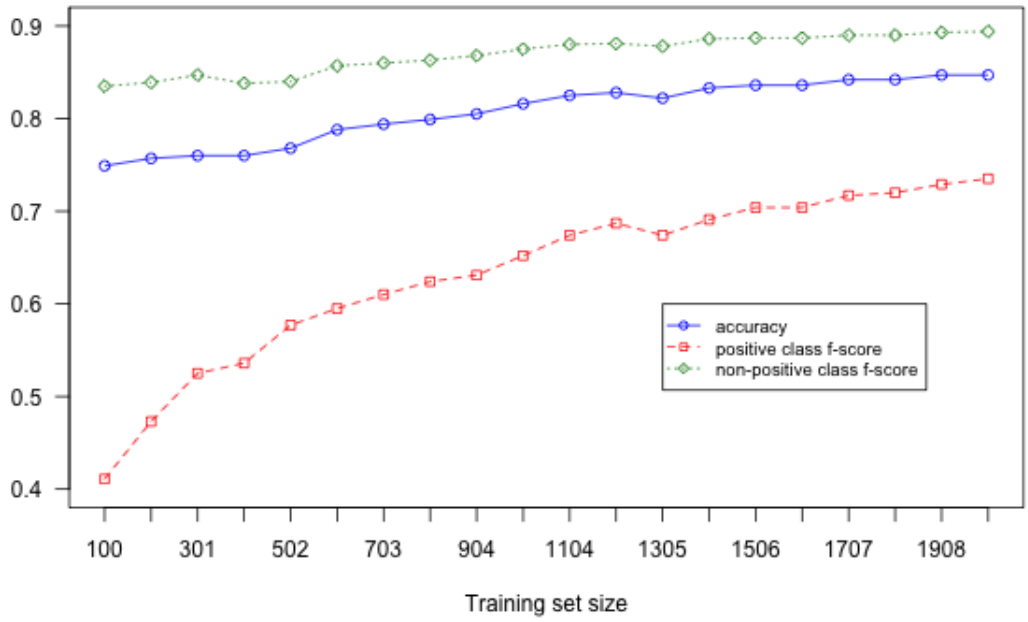


Figure 6.12: Classification accuracies, and positive and non-positive class F-scores for training sets of various sizes.

non-positive, the intervention is not recommended.

3. If the same intervention is mentioned in multiple sentences, the intervention is considered to be a recommendation if any of the sentences associated with it is classified as positive.

Using these rules, and the manually annotated polarities of the 111 interventions as the gold standard, we compute the recall, precision, and F-score of our approach. In this experiment, we apply 10-fold cross validation using the same classifier as above to classify the sentences. To enable a comparison, we introduce three simple baseline systems and compute their recall, precision, and F-scores as well. For the first baseline (B1), we run MetaMap on the key sentences to automatically identify the contexts. All text elements tagged by MetaMap to belong to semantic types representing possible interventions are considered to be recommendations.[12] This simple baseline primarily gives us an idea of the ability of MetaMap to automatically identify interventions for recommendation. For the second baseline (B2), the same automatic approach is used to identify the interventions, and the set of cue words (*good* words) used for feature generation in the classification experiments is used to determine recommendations. If a sentence

[12]The semantic types are: aapp, antb, hops, horm, nnon, orch, phsu, strd, vita, bacs, carb, eico, elii , enzy, imft, inch, lipd, nsba, opco.

containing an intervention also contains a *good* word, the intervention is recommended, otherwise it is not. For the third baseline (B3), the same approach for intervention identification and the same cue words are used, but an intervention is considered to be a recommendation if a *good* word appears in its dependency chain. We define recall and precision for this task as follows:

$$Recall = \frac{CS}{GS} \tag{6.3}$$

$$Precision = \frac{CS}{TS} \tag{6.4}$$

where *CS* represents the number of correct recommendations made by the automatic system, *GS* represents the number of recommendations in the gold standard (111), and *TS* represents the total number of recommendations made by the system. Table 6.5 presents the results of this evaluation. The table shows that for all three baseline systems, the F-scores are very low and comparable. Among the baseline systems, B1 gives the highest recall, but the value is still quite low. For the other two baseline systems, recall is very low indicating that, only in approximately 25% cases, an automatically identified intervention and a cue phrase occurs in the same sentence. Using dependency chains, understandably, decreases recall and increases precision, but very slightly. The results obtained by the baseline systems indicate the difficulty of obtaining the context interventions automatically. The low precisions indicate that even when the relevant contexts are identified, predicting the recommendation using simple rules is difficult. Our approach, using supervised classification, has the advantage of having the contexts manually identified. The F-score obtained is similar to the highest F-score (0.74) obtained for the positive class in our sentence classification task. As explained earlier, due to the lack of training data for this class, the classifier fails to correctly identify the positive polarities in a number of cases, leading to relatively low recall. However, even with the small training set, our approach obtains high precision indicating that the approach has a low possibility of making the wrong recommendation. Having a larger training set is likely to drive the recall/precision values close to the values we obtained via manual annotation.

Discussion

Our investigation of sentence level polarity classification suggests that it is a feasible option for generating bottom-line recommendations. Our approach obtains relatively high accuracies using only a small subset of our corpus as annotated training data. The analysis of the influence of training set size on accuracy indicates that a larger annotated training set will lead to even better

System	Recall	Precision	F-score
B1	0.42	0.27	0.33
B2	0.26	0.44	0.33
B3	0.23	0.47	0.31
Polarity Classification	0.62	0.82	0.71

Table 6.5: Comparison of recall, precision, and F-scores for three baselines and our polarity classification approach.

accuracies. For the positive class, particularly, the gradient of the training set size versus F-score curve remains positive even towards the right of the curve. Thus, it is clear that larger training data will enable the detection of positively polarised key sentences with greater accuracies.

Automatically identifying sentence level polarities is just the first step in the generation of bottom-line summaries, and our experiments simply illustrate the promise of this supervised classification as a possible approach for this task. Generating automatic bottom-line, abstractive summaries is a complex task, and a number of research problems related to sentence level polarity classification must be solved before the whole process can be performed automatically. For example, our approach is dependent on the manual annotation of the contexts. Depending on the types of question, the contexts can be medical concepts such as diagnostic procedures, surgical procedures, pharmacological substances, and so on. In various research tasks mentioned in this thesis, we have relied on MetaMap for the identification of such medical concepts (semantic types). However, the accuracies of MetaMap are not reliable enough to fully automate the identification of contexts. Furthermore, the contexts may also be non UMLS semantic types. Therefore, further research is required to automatically identify the relevant contexts when generating recommendations.

Our analysis also shows that in some cases, the polarities of different sentences relative to the same contexts may not agree. Furthermore, in our analysis, we have ignored factors such as side effects, duration of therapy/management (long term/short term), target populations and so on. Often recommendations proposed by one research may be negated by other research papers. In a fully functioning system, all these *aspects* must be taken into account, and the summaries must indicate recommendations relative to all these aspects. Our work in this chapter is very much the first steps in generating evidence-based, bottom-line recommendations. The success of the polarity classification approach depends on the performances of a number of automatic preprocessing steps, and also future research in the related fields mentioned. We discuss possible future research directions below in this chapter, and also in the next and concluding chapter of the thesis.

6.4 Summary

In this chapter, we have investigated possible approaches for the generation of bottom-line summaries. We first introduced several variants of *coverage scores*, and using them we compared the differences in the extent to which various types of summaries and source abstracts contain the same information as the bottom-line recommendations. We manually analysed the non-covered elements from the different sets of summaries to check how much useful information is lost during the single-document summary generation process. Our analysis showed that human-authored summaries are superior as they contain all the important information in a very compressed manner. In terms of automatic summaries, our summariser (QSpec), described in the previous chapter, is capable of generating content-rich summaries by discarding significant amounts of noise with very little loss in useful content. The Coverage Analysis, described in section 6.2, showed that, although the QSpec summaries had lower coverage compared to the human-authored summaries, most of the missing information was stylistic content and the amount of useful missing information was quite low. Based on this analysis, we argued that there is promise in a two-step summarisation approach, where the first step involves content selection via query-focused, extractive, single-document summarisation, and the second step involves the synthesis of information from the various single-document summaries.

In the second part of the chapter, we explored the applicability of two possible approaches for the generation of bottom-line summaries from single-document summaries. We first explored the possibility of using sentence fusion, which utilises the redundant information in texts from distinct documents, to perform abstractive summarisation. Our experiments showed that the single-document summary sentences exhibit relatively low inter-sentence similarities/redundancies, which makes it difficult to apply redundancy reliant approaches for the task of bottom-line summary generation. Following this, we investigated the possibility of using automatic sentence level polarity classification using supervised machine learning for the task of bottom-line summary generation. Our preliminary analysis, using manually annotated sentences and bottom-line summaries showed that there is very high agreement among the two sets of context-sensitive polarities. This suggests that if sentence level polarities can be automatically classified, they have the potential of being used in bottom-line summary generation. Finally, we experimented with supervised machine learning, using a small sample of manually annotated data and showed that it is possible to obtain high accuracies using a combination of context-sensitive and context-free features. Post-classification analysis also suggest that classification accuracies can be further improved via the use of larger training data.

Our experiments in the second part of this chapter suggest that automatic classification of the polarities of sentences is possible, and this approach can be used for the automatic generation of abstractive bottom-line summaries. However, several related research problems must be solved

before end-to-end systems generating bottom-line recommendations can be implemented. These include the automatic identification of contexts, automatic identification of aspects (*e.g.*, therapy duration, side effects, different modes of treatment, and so on), and combining the polarities of multiple sentences to predict the final recommendations. We leave these as future work and discuss these in more detail in the next chapter.

7 Conclusions

7.1 Thesis Contributions: A Summary

In this thesis, we introduced a model for the generation of automatic summaries for evidence-based medicine. Our research was motivated by the time-associated obstacles hindering the practice of evidence-based medicine. Evidence-based medicine guidelines urge practitioners to obtain the best available medical evidence when answering clinical queries. The purpose of relying on research evidence for decision making is to improve patient care in the long run through the utilisation of patient-oriented evidence. Obtaining the best available evidence, and deriving answers to clinical queries from them, requires practitioners to search for evidence, appraise the quality of the evidence, extract query-relevant evidence, and synthesise evidence from multiple source articles to generate a bottom-line answer. Because of the time consuming nature of this elaborate process, it is often not possible to follow evidence-based medicine guidelines at point-of-care. The possible automation of the various sub-processes of evidence-based answer generation is thus very desirable.

In this thesis, we focused our research on utilising a specialised corpus for performing evidence-based medicine summarisation. We analysed a corpus that was prepared by collecting data from real life evidence-based medicine practice. The corpus enabled us to study the process of evidence-based answer generation in detail and develop approaches specialised to the medical domain. Our study of the corpus revealed that the process of evidence-based answer generation involves several related steps. In particular, the data in our corpus suggests that there may be three tasks involved, following the retrieval of relevant documents associated with a clinical query, to generate bottom-line, evidence-based answers. These three steps are:

1. Extraction of query-relevant information from individual medical publications.
2. Synthesis of extracted information from multiple documents to generate bottom-line

recommendations.

3. Appraisal of evidence and the grading of it on a chosen scale.

We addressed these three tasks in our work. We modelled the process of evidence appraisal as a supervised classification task and applied machine learning algorithms to solve the classification problem. For the task of extracting query-relevant information, we used a sentence level extractive summarisation model. We derived various statistics from the corpus to rank sentences, and the highest ranked sentences were chosen to be in the final summaries. We modelled the task of generating bottom-line recommendations as an abstractive summarisation task, and we applied automatic, sentence level, context-sensitive polarity classification to predict bottom-line recommendations from the single-document, extractive summaries. Since the task of generation of bottom-line recommendations relies on the single-document extracts, we applied a two-step summarisation approach in our model.

The key contribution in this thesis is the presentation of models and algorithms that automate the generation of evidence-based summaries from source texts. In particular, the thesis proposed approaches for utilising domain-specific information in various ways for improving performance on all the three tasks mentioned above. Furthermore, we extensively used statistics from our annotated corpus, and the approaches we proposed are data-driven, rather than intuition oriented. The presence of such specialised data also enabled us to devise automatic evaluation techniques through which we evaluated the performance of various components against gold standards. In the following subsections, we revisit the main characteristics of the work described in this thesis.

7.1.1 A Model for Appraising the Qualities of Evidence

In our corpus, each evidence-based answer is allocated a SORT grade (A, B or C) indicating its quality. In our model, we considered these grades to be classes for the texts associated with the grades, and applied supervised machine learning to automatically predict these grades. The first step in our task was to identify important features that influence the qualities of evidence. We discovered that the publication types of individual articles are useful predictors of evidence grades, and the titles of articles are also useful. However, publication dates (years) and publication venues (*i.e.*, journal names) of individual articles are not useful predictors of evidence grades according to our analysis.

Due to the importance of the information regarding the publication types of individual articles in the grade classification process, we attempted to devise an automatic approach for identifying the publication types of medical articles. We showed that a rule-based approach can efficiently identify the publication types of high quality articles such as Systematic Reviews and Randomised

Controlled Trials by utilising information from the article titles, abstracts, and the associated meta-data. Automatic identification of lower quality publication types such as Case Studies is more challenging since the article titles and abstracts often do not contain the necessary information.

We applied supervised machine learning with automatically extracted features to perform the grading task. We applied a sequence of classifiers that attempted to separate A and C grade evidences from B grade ones. We obtained an accuracy of over 60% using this approach, which was a significant improvement over the baseline. Our system participated in the 2011 ALTA shared task [Mollá and Sarker, 2011] and obtained the top position, which illustrates the validity and quality of this approach. We introduced an evaluation metric called Average Error Distance (AED), which attempts to estimate the *closeness* of a system's grades to actual grades, and we showed that our sequential classification model achieves improved AEDs compared to the baseline.

To conclude our research on this topic, we conducted a human evaluation and compared the performance of our system with human experts. Our experiments revealed that when human experts are given the same data as our machine learning algorithm, they only have *moderate* agreement regarding the grades. The experiments also revealed that, although the performance of the experts is comparable to our system, when compared against the gold standard, there are still significant disagreements between the expert assigned grades, and the grades assigned by our system. Based on our findings, we can conclude that supervised classification is a promising approach for automatic grade recommendations. Our evaluations suggested that it may not be possible for an evidence grading system to *markedly* improve on the performance using the data that is currently available because of the relatively low human-human agreements. Our evaluations also showed that human-human agreements are higher than human-system agreements. This suggests that *minor* improvements can perhaps be made to the system, using the current data, to increase its agreement with the human experts.

7.1.2 A Model for Content Selection via Query-focused, Extractive Summarisation

We applied a simple, data-driven sentence extraction model for the summarisation task. We divided our specialised corpus into two parts — one for obtaining statistics and the other for evaluation. From the source documents in the training set, we generated the best three sentence summaries. We then used these best sentences, and the human generated summaries associated with each source document, to compute various statistics which we used for summarisation. We used features such as relative sentence positions, sentence lengths, the PIBOSO classifications of sentences, similarities between sentences and the associated queries, the semantic types present

in sentences, and the semantic associations between sentences and the associated queries. For the sentence classification problem, we designed a classifier that obtained second position in the 2012 ALTA shared task[1]. We showed that using carefully extracted statistics from a specialised corpus significantly improves summarisation performance. We applied a strategy, which we call target-sentence-specific summarisation. Using this strategy, we applied different statistics for different target summary sentences. We also modified the Maximal Marginal Relevance approach to enable concept matches, and significantly utilised available domain knowledge. Furthermore, we showed that customising the summarisation technique to the type of question improves summarisation performance.

We used the automatic summary evaluation tool ROUGE to evaluate our extractive summaries relative to the human generated summaries. We compared our summarisation system to various established baselines for this domain using a percentile-rank based approach. The best ROUGE-L F-score obtained by our system is a statistically significant improvement over the best performing baseline system, and has a percentile rank of 96.8%.

7.1.3 A Model for the Generation of Bottom-line Recommendations from Single-document Extracts

We explored possible approaches for the generation of bottom-line recommendations, and the possibility of using single-document summaries, rather than full source abstracts, to generate the final recommendations. For our investigation, we introduced several variants of *coverage scores*, and using them we compared the differences in the extent to which various types of summaries and source abstracts contain the information from the bottom-line recommendations. Our analysis showed that our single-document summariser (QSpec) is capable of generating content-rich summaries by discarding significant amounts of noise. Our scrutiny of the *uncovered* elements from this analysis showed that only a small percentage of useful content is lost during the summarisation task by QSpec. Based on this analysis, we argued that there is promise in a two-step summarisation approach, where the first step involves content selection via query-focused, extractive, single-document summarisation, and the second step involves the synthesis of information from the various single-document summaries.

We investigated the applicability of two possible approaches for the generation of bottom-line summaries from single-document summaries. We first explored the possibility of using sentence fusion, which utilises the redundant information in texts from distinct documents, to perform abstractive summarisation. Our experiments showed that the single-document summary sentences exhibit relatively low inter-sentence similarities/redundancies, which makes it difficult to apply

[1] http://www.alta.asn.au/events/sharedtask2012/. Accessed on 26th May, 2014.

redundancy reliant approaches for the task of bottom-line summary generation. Following this, we investigated the possibility of using automatic sentence level polarity classification using supervised machine learning for the task of bottom-line summary generation. We experimented with supervised machine learning, using a small sample of manually annotated data and showed that it is possible to obtain high accuracies using a combination of context-sensitive and context-free features. Our post-classification analysis suggested that classification accuracies can be further improved via the use of larger training data.

7.2 Future Work

In this section, we briefly discuss possible future work that we envision based on the research described in this thesis.

7.2.1 Automatic Retrieval of Relevant Documents

In our work, we provide our system with the relevant documents. Thus, our system operates under the assumption that an IR process identifies the relevant documents and filters out unimportant documents. For future work, it will be interesting to integrate an IR system and explore the impact of using automatically retrieved documents as inputs to our system. It is likely that this approach of selecting documents will introduce noise, negatively affecting the performance of our system. However, they may also introduce redundancies that can be exploited to build the final summaries.

7.2.2 Improving Automatic Evidence Grading

Our experiments on evidence grading verified that a supervised classification model can be effectively applied to automate the process. Our evaluations revealed that the performance of our approach is comparable to human performance on the same data (when compared against the gold standard). However, we also discovered that inter-human agreements for the grades are higher than system-human agreements. This suggests that there is room for modifications to the approach, which can increase the system-human agreements. In our supervised classification strategy, we incorporated publication types, word n-grams and title n-grams as features. When human experts perform the grading, they take into account more factors. These include, among other things, the sizes of studies, the type of evidence (*i.e.*, patient-oriented/disease-oriented), and consistency of outcomes between distinct studies. Solving each of these problems individually are interesting tasks for the future. It is possible that if each of these problems can be solved, similar to our approach to identify the publication types of medical papers, the outputs of these

processes can be plugged into our classification algorithm. However, these are also difficult problems to solve. We have attempted to incorporate consistency information by identifying document-level polarity classification [Sarker et al., 2011], but the results are not sufficiently good to utilise that approach as an intermediate step in the evidence grading task. Attempting to solve these intermediate problems and incorporating the results of these tasks in the evidence grading approach will be very interesting future challenges to pursue.

7.2.3 Improving Content Extraction from Single Documents

Our extractive summarisation approach showed excellent performance compared to baseline and benchmark systems. There is, however, still room for improvement. We used various intermediate steps to generate the features for our summarisation task. These include automatic sentence classification and question classification. It will be interesting future research to investigate how the performance of our summarisation module is affected if improvements in these intermediate steps can be achieved.

We used a medium sized corpus for our research. Thus, for some of the features, there was little data available for the generation of statistics. For example, in our *query type* dependent scores, we generated statistics for each question type. For some of the question types, such as *History* and *Device*, our corpus only contained a few samples. Having a larger corpus would make the statistics associated with sparse data more reliable. Thus, future research should focus on the generation of more annotated data, such that sufficient data is available for the generation of the various statistics.

Our summariser generates three sentence extracts. There are, however, cases when three sentences are not sufficient to cover all the information required to generate bottom-line recommendations. Thus, long term future research should investigate the possibility of generating single-document abstractive summaries that combine required information from more than just three sentences.

7.2.4 Bottom-line Summary Generation

Our experiments on predicting bottom-line recommendations from single-document extracts produced promising results, suggesting that context-sensitive polarity classification is a possible approach for the generation of bottom-line recommendations. However, applying this approach in real life evidence-based medicine practice requires solving several related problems. We now briefly discuss some possible future work in this area.

Annotation of Data

In Chapter 6, our sentence level polarity classification experiments clearly indicated that the availability of large amounts of annotated data will increase the performance of the classification strategy. Thus, the most important future task that we would like to undertake is the annotation of large volumes of data which can be utilised in the supervised classification task.

Automatic Identification of Context Interventions

In our context-sensitive polarity classification approach, we manually specified the context-interventions. In a fully automatic recommendation generation system, the interventions need to be automatically identified. Our brief analysis suggested that automatic identification of the context interventions is a difficult task. So, it will be very important to focus future research on investigating techniques for the automatic identification of candidate context interventions from sentences.

Cross-sentence polarities

Finally, a small number of mistakes in the recommendations generated by our system were due to conflicting information presented in distinct sentences. Future research should focus on investigating approaches for combining multiple sentence level polarities associated with the same context interventions.

7.3 Final Words

This thesis described computational models for the core aspects of automatic evidence-based text summarisation. Our research identified the key tasks associated with evidence-based answer generation and proposed models for two crucial tasks: automatic appraisal of evidence and automatic bottom-line recommendation generation. The new insights gained from this research take us closer towards a fully automated Question Answering system that can take natural language questions as input and generate bottom-line recommendations in response.

A Sample PubMed Abstract

```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle,
1st January 2011//EN"
"http://www.ncbi.nlm.nih.gov/entrez/query/DTD/pubmed_110101.dtd">
<pubmedarticleset>
 <pubmedarticle>
  <medlinecitation owner="NLM" status="MEDLINE">
   <pmid version="1">
    8600478
   </pmid>
   <datecreated>
    <year>
     1996
    </year>
    <month>
     05
    </month>
    <day>
     01
    </day>
   </datecreated>
   <datecompleted>
    <year>
     1996
    </year>
    <month>
```

```
    05
   </month>
   <day>
    01
   </day>
  </datecompleted>
  <daterevised>
   <year>
    2009
   </year>
   <month>
    11
   </month>
   <day>
    19
   </day>
  </daterevised>
  <article pubmodel="Print">
   <journal>
    <issn issntype="Print">
     0033-3174
    </issn>
    <journalissue citedmedium="Print">
     <volume>
      57
     </volume>
     <issue>
      6
     </issue>
     <pubdate>
      <medlinedate>
       1995 Nov-Dec
      </medlinedate>
     </pubdate>
    </journalissue>
    <title>
     Psychosomatic medicine
    </title>
```

<isoabbreviation>
Psychosom Med
</isoabbreviation>
</journal>
<articletitle>
Nonpharmacological management of headaches during pregnancy.
</articletitle>
<pagination>
<medlinepgn>
527-35
</medlinepgn>
</pagination>
<abstract>
<abstracttext>

Concerns about the effects of maternal medications on the growing baby limit the use of medication treatment for benign conditions, such as recurring headaches, during pregnancy and lactation. Nonpharmacological therapies hold particular promise for pregnant women due to the limited medication options. No controlled studies, however, have reported on the efficacy of nonpharmacological treatments for pregnant women. The first study evaluated the effectiveness of a combined nonpharmacological treatment (CT) consisting of relaxation, skin-warming biofeedback, and physical therapy for pregnant women with chronic headaches. In a second study, the CT protocol was compared with an attention control (AC) that received headache education and skin-cooling biofeedback. The first study resulted in significant symptom improvement in 79% of subjects, with an overall 72.9% reduction in headaches. In the second study, both groups improved with treatment; however the CT group was more likely to experience significant headache relief (72.7%) than the AC group (28.6%, chi 2(1) = 4.97, p < .03). Significant improvement was maintained at a 6-month follow-up for over 50% of patients. It is concluded that the combined nonpharmacological treatment was more effective than an attention control in reducing headaches during pregnancy. This treatment was effective regardless of predisposing variables.

</abstracttext>
</abstract>

B Important Question and Answer Semantic Types

Here we present the important question and answer semantic types, for each of the twelve categories of questions, that we identified during our analysis explained in Section 5.4.3. These semantic types are utilised when performing sentence scoring using semantic associations.

Treatment and Prevention:
Important question semantic types: topp, dsyn, phsu, hlca, orch, strd, sosy, mobd, bacs, mobd, inpo, fndg, ftcn, acab, anab
Important answer semantic types: phsu, topp, orch, acty, gngm, aggp, sosy, hlca, resa, ftcn, inpr, podg, spco, idcn, popg, dsyn, fndg

Diagnosis:
Important question semantic types: diap, dsyn, fndg, patf, lbpr, medd, hlca, ftcn, sosy, mnob, inpo, idcn, orch, bsoj, blor, bdsy, bpoc, bdsu
Important answer semantic types: diap, lbpr, patf, dsyn, clna, fndg, bpoc, hlca, sosy, ftcn, popg, spco, inpr, aggp, resa, acty, topp, phsu, orch

History:
Important question semantic types: ocdi, imft, aapp, sosy, podg, popg, sosy
Important answer semantic types: irda, menp, patf, bodm, sosy, cnce, dsyn, podg, clna, idcn, acty, ftcn, diap, lbpr, resa, fndg, orch, topp, phsu

Etiology:
Important question semantic types: bdsu, cnce, bmod, socb, lbpr, antb, mobd, sosy, orch, hlca, ftcn, dsyn

Important answer semantic types: antb, inpr, aapp, orch, dsyn, patf, cnce, lbpr, hlca, aggp, popg, phsu, podg, ftcn, fndg, acty

Pharmacological:
Important question semantic types: antb, strd, orch, phsu, bacs, podg, patf, dsyn, bpoc, orgf, popg, topp, ftcn, sosy, aggp, fndg
Important answer semantic types: antb, orch, phsu, resa, podg, ftcn, inpr, idcn, topp, spco, aggp, dsyn, popg, fndg, strd, bacs

Prognosis:
Important question semantic types: bmod, strd, podg, lbpr, spco, orch, fndg, inpr, ftcn, phsu, idcn, dsyn, acty, topp, hlca
Important answer semantic types: acty, bmod, menp, podg, lbpr, hlca, ftcn, fndg, dsyn, qlco, orch, idcn, phsu, topp, sosy, aggp

Management:
Important question semantic types: orga, carb, ocac, cnce, spco, dora, inpo, fndg, idcn, bpoc, sosy, orgf, aggp, podg, menp, ftcn, dsyn, orch, patf, hlca, popg, topp, phsu
Important answer semantic types: diap, gngm, bpoc, fndg, dsyn, podg, hlca, idcn, aggp, inpr, ftcn, sosy, popg, acty, resa, topp, orch

Physical Finding:
Important question semantic types: clna, prog, vita, virs, orga, blor, patf, inch, cnce, diap, sosy, socb, orgf, hlca, food, orch, menp, topp, popg, ftcn, dsyn, podg, phsu
Important answer semantic types: aggp, lbpr, prog, orgf, orga, horm, patf, fndg, diap, idcn, hlca, inpr, sosy, popg, topp, dsyn, ftcn, acty, resa, podg

Epidemiology:
Important question semantic types: carb, virs, orga, bmod, hlca, socb, neop, inpr, bacs, strd, menp, sosy, idcn, podg, dsyn, popg, fndg, phsu, aggp, ftcn
Important answer semantic types: carb, inpo, sosy, orgf, horm, orga, podg, dsyn, bpoc, hlca, diap, inpr, fndg, lbpr, fndg, lbpr, ftcn, idcn, phsu, acty, topp, orch

Test:

Important question semantic types: lbpr, neop, hlca, orga, diap, socb, medd, mnob, patf, bacs, idcn, aggp, fndg, bpoc, dsyn, ftcn, popg, sosy, podg, phsu
Important answer semantic types: lbpr, diap, hlca, clna, dsyn, fndg, bpoc, popg, gngm, orga, ftcn, podg, aggp, idcn, inpr, resa, acty, topp, phsu, orch, neop

Device:
Important question semantic types: medd, phpr, mnob, dsyn
Important answer semantic types: phpr, medd, antb, patf , blor, popg, topp, fndg, bpoc, hlca, orch, resa, inpr, dsyn

Procedure:
Important question semantic types: acty, bmod, bodm, inpo, mnob, orga, blor, inbe, inch, neop, topp, patf, bpoc, idcn, diap, medd, popg, aggp, strd, orgf, hlca, lbpr, podg, fndg, dsyn, orch, ftcn, resa, hlca, inpr
Important answer semantic types: bmod, neop, orga, parf, diap, mnob, fndg, topp, dsyn, podg, popg, spco, bpoc, acty, idcn, ftcn, resa, hlca, inpr, phsu

C Sample Single-document Summaries

Question: Can nonantidepressants help treat depression?

QSpec summary: An attempt was made to identify all placebo-controlled trials of lithium augmentation in refractory depression. Aggregating three studies with a total of 110 patients that used a minimum lithium dose of 800 mg/day, or a dose sufficient to reach lithium serum levels of > or = 0.5 mEq/L, and a minimum treatment duration of 2 weeks, the authors found that the pooled odds ratio of response during lithium augmentation compared with the response during placebo treatment was 3.31 (95% confidence interval, 1.46-7.53). The authors conclude from this meta-analysis that with respect to efficacy, lithium augmentation is the first-choice treatment procedure for depressed patients who fail to respond to antidepressant monotherapy.

Human authored summary: A meta-analysis concluded that a lithium dose sufficient to produce serum levels of at least 0.5 mEq/L and a minimum treatment duration of 2 weeks resulted in a pooled OR of response to lithium augmentation compared with placebo of 3.31 (95% CI, 1.46-7.53). Lithium augmentation is a reasonable alternative for depressed patients who don't respond to conventional antidepressants.

PubMed ID: 10505584

Question: Do topical antibiotics improve wound healing?

QSpec summary: Triple antibiotic ointment, containing bacitracin, polysporin, and neomycin, was compared to placebo ointment. Infection occurred significantly more often in children using placebo ointment than in those using topical antibiotic (47% vs. 15%; p = 0.01). This study further confirms the importance of skin carriage of group A streptococci as a precursor to pyoderma and demonstrates the importance of minor skin trauma as a predisposing factor.

Human authored summary: A double-blind study of 59 patients found Neosporin superior to placebo ointment in the prevention of streptococcal pyoderma for children with minor wounds. Infection occurred in 47% of placebo-treated children compared with 15% treated with the triple-antibiotic ointment (NNT=32; P=.01).

PubMed ID: 2995463

Question: Do topical antibiotics improve wound healing?

QSpec summary: To evaluate the ability of a novel topical antimicrobial gel containing cetrimide, bacitracin, and polymyxin B sulfate to prevent infections of minor wounds. Accidental injuries occurring at school were treated in a standardized manner by nurses at each site. The novel gel preparation containing cetrimide, bacitracin, and polymyxin B sulfate showed therapeutic action and reduced the incidence of clinical infections in minor accidental wounds.

Human authored summary: A clinical trial compared the efficacy of a cetrimide, bacitracin zinc, and polymyxin B sulfate gel (a combination not available in the US) with placebo and povidone-iodine cream in preventing infections in 177 minor wounds (cuts, grazes, scrapes, and scratches) among children. The antibiotic gel was found to be superior to placebo and equivalent to povidone-iodine, in that it reduced clinical infections from 12.5% to 1.6% (absolute risk reduction [ARR]=0.109; 95% confidence interval [CI], 0.011-0.207; NNT=11).

PubMed ID: 9161648

Question: Do topical antibiotics improve wound healing?

QSpec summary: We compared mupirocin cream with oral cephalexin in the treatment of wounds such as small lacerations, abrasions, or sutured wounds. In 2 identical randomized double-blind studies, 706 patients with secondarily infected wounds (small lacerations, abrasions, or sutured wounds) received either mupirocin cream topically 3 times daily or cephalexin orally 4 times daily for 10 days. The occurrence of adverse experiences related to study treatment was similar for the 2 groups, with fewer patients in the mupirocin cream group reporting diarrhea (1.1% vs 2.3% for cephalexin).

Human authored summary: A study with 2 parallel, identical RCTs of a total of 706 patients found mupirocin cream (Bactroban) to be equivalent to oral cephalexin in the treatment of secondarily infected minor wounds, such as small lacerations, abrasions, or sutured wounds. Clinical success (95.1% for mupirocin and 95.3% for cephalexin), bacteriologic success (96.9% for mupirocin and 98.9% for cephalexin), as well as the intention-to-treat success rate of 83% at follow-up were equivalent in the 2 groups.

PubMed ID: 9866667

Question: Should you use steroids to treat infectious mononucleosis?

QSpec summary: To evaluate the efficacy of a single oral dose of dexamethasone for pain relief in acute exudative pharyngitis associated with infectious mononucleosis. Patients aged between 8 and 18 years with a sore throat from clinically suspected infectious mononucleosis were eligible. The short-lived relief of pain in acute exudative pharyngitis in children with suspected infectious mononucleosis may suggest that a single oral dose of dexamethasone may not be sufficient and that additional doses may be necessary for ensuring lasting relief.

Human authored summary: In an RCT of 40 patients, 1 dose of dexamethasone reduced throat pain at 12 hours in 60% of the treatment group, compared with placebo. However, no significant differences were noted at 1 and 7 days.

PubMed ID: 14993084

Question: How effective are pharmacologic agents for alcoholism?

QSpec summary: Nalmefene is a newer opioid antagonist that is structurally similar to naltrexone but with a number of potential pharmacological advantages for the treatment of alcohol dependence, including no dose-dependent association with toxic effects to the liver, greater oral bioavailability, longer duration of antagonist action, and more competitive binding with opioid receptor subtypes that are thought to reinforce drinking. Significantly fewer patients treated with nalmefene than patients given placebo relapsed to heavy drinking through 12 weeks of treatment (P<.02), with a significant treatment effect at the first weekly study visit (P<.02). Treatment with nalmefene was effective in preventing relapse to heavy drinking relative to placebo in alcohol-dependent outpatients and was accompanied by acceptable side effects.

Human authored summary: Naltrexone (50 mg qd), nalmefene (10-80 mg qd), and acamprosate (dose based on patient weight) are all superior to placebo and other agents such as the SSRIs, disulfiram, and serotonergic agents in reducing relapse rates and the phenomena of craving and in increasing abstinence rates.

PubMed ID: 10435606

Question: What medications are effective for treating symptoms of premenstrual syndrome (PMS)?

QSpec summary: Forty women with premenstrual tension received either placebo, 100, 200 or 400 mg danazol daily for 3 months in a pilot study arranged as a double-blind trial. In patients treated with danazol, symptom scores for breast pain during the second and third months and for irritability, anxiety and lethargy during the third month were significantly (P less than 0.05) lower than scores in those given placebo. By the end of the trial more than 75% of patients who were still taking danazol were essentially free of breast pain, lethargy, anxiety and increased appetite, but results for other common symptoms were no better than with placebo.

Human authored summary: Gonadotropin-releasing hormone agonists may be effective, but troublesome anti-estrogenic side effects limit their utility. Estrogen and progesterone “add-back" therapy to counter side effects further complicates this approach. The gonadotropin inhibitor danazol has a high treatment dropout rate at higher doses 200-400 mg/d continuously, but can be effective in individuals who are able to tolerate it.

PubMed ID: 3545282

D List of Words for the Change Phrases Features

This list shows the list of **good**, **bad**, **more** and **less** words that were used for the *change phrases* features.

Good:
benefit, beneficial, improve, advantage, resolve, good, fantastic, relief, superior, efficacious, effective, improve effectiveness, importance of protecting, significant advantage, significant therapeutic advantage, may be effective, effective approach, simple and effective, simple and effective treatment, safe, well tolerated, well-tolerated, useful, maybe useful, illustrate the benefits, significant improvement, significantly improve, clinically worthwhile, worthwhile, recover rapid, satisfactory outcome, satisfactory, similarly effective, supports, approve, more effective, high efficacy, cured, vitality, relaxing, benefit, tolerability, improvement right, effective, stable, best, better, pleasurable, relaxation, favour, beneficial, safety, prevents, successful satisfaction, significant, superior, contributions, reliability, robust, tolerated, improving, survival, favourable, reliable, recovered, judiciously, consciousness, efficacy, prevented, satisfied, prevent, advantage, encouraging, tolerance, success, significance, improve, improvements

Bad:
complication, risk, adverse, mortality, morbidity, death, fatal, danger, no benefit, discourage, short-term risk, long-term risk, damage, ineffective, suffer, depression, acute, sore, outpatient, disabling, diabetes, difficulties, dysfunction, distorted, poorer, unable, prolonged, irritation, disruptive, pathological, mutations, disease, infection, harms, difficulty, weakened, inactive, stressors, hypertension, adverse, insomnia, relapsing, malignant, suffer, exacerbate, dryness, fever, overestimate, constipation, deposition, colic, tension, hazards, diarrhoea, weakness, irritability, insidious, distress, weak, cancer, emergency, risk, block, unsatisfactory, blinding, nausea, traumatic, wound, intention, loses, intensive, relapse, recurrent, extension, die, cancers, malaise, crying,

toxic, injury, confounding, complaints, misuse, insignificant, poisoning, anoxic, amputation, death, nightmares, deteriorate, fatal, injuries, fatigue, invasive, suicide, chronic, relapsed, disturbances, confusion, died, fluctuating, severities, delusions, compulsions, conflict, trauma, cried, impair, severe, tremor, weaker, illness, inpatients, worry, rebound, worse, reversible, dizziness, attacks, pointless, disorders, dyskinesia, risks, fatty, negative, conflicting, upset, fishy, hard, harm, bleeding, inflammatory, hampered, underpowered, obstruction, headache, problem, bleeds, panic, loss, odds, retardation, dysfunctional, render, difficult, drowsiness, lack, suicidal, obsessions, impaired, cough, severity, suffering, violent, strokes, virus, stroke, flatulence, fibrates, blind, burning, faintness, suffered, threatening, misdiagnosing, bitter, excessive, diabetics, malfunction, abnormal, deterioration, bad, confounded, sadness, mortality, disturbance, agitated, attack, infections, negativistic, deaths, poor, wrong, worsening, adversely, insufficient, scarring, headaches, disability, overdose, serious, delayed, discomfort, sweating, morbidity, nerve, parkinson, toxicity, nervous, pain, stress, weakens, incorrect, disorder, worsened, malformations, blinded, rigidity, prolong, adversity, abuse, lacked, dyspepsia, sads, onset, failure, inadequate, sensitivity, impairment, dementia, harmful

More:
enhance, augment, increase, amplify, raise, boost, add to, higher, exceed, rise, go up, surpass, more, additional, extra, added, greater, positive, high, prolong, increase, enhance, elevation, higher, exceed, enhancement, peaked, more, excess

Less:
drop, fewer, slump, fall, down, pummel, less, lower, low, decrease, reduce, decline, descend, collapse, fail, subside, lesser, poorer, worse, smaller, negative, prevent, reduce, prevent, below, lower, decrease, fall, low, reduce, decline, less, little, mild, drop, fewer

Bibliography

Stergos D. Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. Summarization from medical documents: A survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

Ramiz M. Alyguliyev. Automatic Document Summarization by Sentence Extraction. *Journal of Computational Technologies*, 12:5–15, 2007.

Rie Kubota Ando, Branimir K. Boguraev, Roy J. Byrd, and Mary S. Neff. Multi-document Summarization by Visualizing Topical Content. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, pages 79–88, 2000.

Okurowski C. Aone. *Advances in Automatic Text Summarization*, chapter A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques, pages 71–80. MIT Press, 1999.

Yindalon Aphinyanaphongs, Ioannis Tsamardinos, Alexander Statnikov, Douglas Hardin, and Constantin F. Aliferis. Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association (JAMIA)*, 12(2):207–216, 2005.

E. C. Armstrong. The well-built clinical question: the key to finding the best evidence efficiently. *Wisconsin Medical Journal*, 98(2):25–28, 1999.

Alan R. Aronson. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 17–21, 2001.

Sofia J. Athenikos and Hyoil Han. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24, 2009.

David Atkins, Dana Best, Peter A. Briss, Martin Eccles, Yngve Falck-Ytter, Signe Flottorp, Gordon H. Guyatt, Robin T. Harbour, Margaret C. Haugh, David Henry, Suzanne Hill, Roman Jaeschke, Gillian Leng, Alessandro Liberati, Nicola Magrini, James Mason, Philippa Middleton, Jacek Mrukowicz, Dianne O'Connell, Andrew D. Oxman, Bob Phillips, Holger J. Schunemann, Tessa Tan-Torres Edejer, Helena Varonen, Gunn E. Vist, John W. Williams,

Stephanie Zaza, and G. R. A. D. E. Working Group. Grading quality of evidence and strength of recommendations. *BMJ*, 328(7454):1490–1497, June 2004.

G. Octo Barnett, Richard N. Winickoff, Mary M. Morgan, and Rita D. Zielstorff. A Computer-Based Monitoring System for Follow-Up of Elevated Blood Pressure. *Medical Care*, 21(4): 400–409, 1983.

Henry C. Barry, Mark H. Ebell, Allen F. Shaughnessy, David C. Slawson, and Fern Nietzke. Family Physicians' Use of Medical Abstracts To Guide Decision Making: Style or Substance? *The Journal of the American Board of Family Practice*, 14(6):437–442, 2001.

Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, 1997.

Regina Barzilay and Kathleen McKeown. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–328, 2005.

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 550–557, 1999.

P. B. Baxendale. Machine-made index for technical literature - an experiment. *IBM Journal of Research Development*, 2(4):354–361, 1958.

Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2(5):S4, 2011.

Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32(Database-Issue):267–270, 2004.

Branimir Boguraev, Rachel Bellamy, and Calvin Stwart. Summarisation miniaturisation: Delivery of news to hand-helds. In *Proceedings of the NAACL-01 Workshop on Automatic Text Summarization*, pages 99–108, 2001.

Andrew Booth, Alan J. O'Rourke, and Nigel J. Ford. Structuring the pre-search reference interview: a useful technique for handling clinical questions. *Bulletin of the Medical Library Association*, 88(3):239–246, July 2000.

Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–686, 1995.

Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont D. Antieau, Andrew Bennett, James J. Cimino, John W. Ely, and Hong Yu. AskHermes: An Online Question Answering System for Complex Clinical Querstions. *Journal of Biomedical Informatics*, 44(2):277 – 288, 2011.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

Hakan Ceylan, Rada Mihalcea, Umut Özertem, Elena Lloret, and Manuel Palomar. Quantifying the limits and success of extractive summarization systems across domains. In *Proceedings of NAACL*, pages 903–911, 2010.

Soumen Chakrabarti, Mukul Joshi, and Vivek Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 208–216, 2001.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. Evaluation of negation phrases in narrative clinical reports. *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 105–109, 2001.

Grace Y. Cheng. A study of clinical questions posed by hospital clinicians. *Journal of the Medical Library Association*, 92(3):445–458, 2004.

Michael G. Christel, Alexander G. Hauptmann, Howard D. Wactlar, and Tobun D. Ng. Collages as dynamic summaries for news video. In *Digital Video Summaries, Idexing and Retrieval*, pages 561–569. ACM Press, 2002.

Jeffrey A. Claridge and Timothy C. Fabian. History and development of evidence-based medicine. *World Journal of Surgery*, 29(5):547–553, May 2005.

Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 358–365, 2001.

Paul Compton and Bob Jansen. Knowledge in Context: A Strategy for Expert System Maintenance. In *Proceedings of the 2nd Australian Joint Artificial Intelligence Conference*, pages 292–306, 1988.

John M. Conroy and Dianne P. O'Leary. Text Summarization via Hidden Markov Models. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 406–407, 2001.

Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, October 1992.

Herma Coumou and Frans Meijman. How do primary care physicians seek answers to clinical questions? A literature review? *Journal of the Medical Library Association*, 94(1):55–60, 2006.

David G. Covell, Gwen C. Uman, and Phil R. Manning. Information Needs in Office Practice: Are They Being Met? *Annals of Internal Medicine*, 103(4):596 – 599, 1985.

James M. Crawford. Original research in pathology: judgement, or evidence-based medicine. *Laboratory Investigation*, 87(2):104–114, 2007.

Laurie Damianos, Steve Wohlever, Jay Ponte, George Wilson, Florence Reeder, Tom McEntee, Robyn Kozierok, Lynette Hirschman, and David Day. Real users, real data, real problems: the MiTAP system for monitoring bio events. In *Proceedings of the second international conference on Human Language Technology Research*, pages 357–362, 2002.

Dipanjan Das and Andre F. Martins. A Survey on Automatic Text Summarization. *Carnegie Mellon University*, 2007.

Frank Davidoff, David L. Sackett, Brian Haynes, and Richard Smith. Evidence based medicine. *BMJ*, 310:1085–1086, 1995.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. Methodological Review: What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, October 2009.

Josephine L. Dorsch. Information needs of rural health professionals: a literature review. *Bulletin of the Medical Library Association*, 88(4):346–354, 2000.

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. Web question answering: is more always better? In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 291–298, 2002.

L. L. Earl. Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, 6:313–334, 1970.

Mark H. Ebell, Jay Siwek, Barry D. Weiss, Steven H. Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *American Family Physician*, 69(3): 548–556, February 2004.

H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285, 1969.

Noémie Elhadad. *User-Sensitive Text Summarization: Application to the Medical Domain*. PhD thesis, Columbia University, 2006.

Noémie Elhadad and Kathleen R. McKeown. Towards generating patient specific summaries of medical articles. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*, pages 31–39, 2001.

Noémie Elhadad, Min-Yen Kan, Judith L. Klavans, and Kathleen McKeown. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*, 33 (2):179–198, 2005.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, George R. Bergus, Barcey T. Levy, M. Lee Chambliss, and Eric R. Evans. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361, August 1999.

John W. Ely, Jerome Osheroff, Paul Gorman, Mark Ebell, Lee Chambliss, Eric Pifer, and Zoe Stavri. A taxonomy of generic clinical questions: classification study. *BMJ*, 321:429–432, 2000.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, M. Lee Chambliss, D. C. Vinson, James J. Stevermer, and Eric A. Pifer. Obstacles to answering doctors' questions about patient care with evidence: Qualitative study. *BMJ*, 324(7339):710, 2002.

John W. Ely, Jerome A. Osheroff, M. Lee Chambliss, Mark H Ebell, and Marcy E. Rosenbaum. Answering physicians' clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association (JAMIA)*, 12(2):217–224, 2005.

Brigitte Endress-Niggemeyer, Jerry Hobbs, and Karen Sparck Jones. Summarising text for intelligent communication. Technical report, Dagstuhl-Seminar-Report, Germany: IBFI GmbH Schloss Dagstuhl, 1995.

Gunes Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

R. Evans, R. Gaizauskas, L. J. Cahill, J. Walker, J. Richardson, and A. Dixon. POETIC: A System for Gathering and Disseminating Traffic Information. *Natural Language Engineering*, 1(4):363–387, 1995.

Atefeh Farzindar and Guy Lapalme. Legal text summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out Conference held in conjunction with ACL 2004*, pages 27–38, 2004.

Mohamed Abdel Fattah and Fuji Ren. GA, MR, FFNN, PNN and GMM based Models for Automatic Text Summarization. *Computer Speech and Language*, 23:126–144, 2009.

Elena Filatova and Vasileios Hatzivassiloglou. Event-based extractive summarization. In *Proceedings of the Second Baltic Conference on Human Language Technologies, Talinn*, pages 104–111, 2004.

Seeger Fisher and Brian Roark. Query-focused summarization by supervised sentence ranking and skewed word distributions. In *Proceedings of the Document Understanding Workshop (DUC 2006)*, pages 8–15, June, 2006.

Marcelo Fiszman, Thomas C. Rindflesch, and Halil Kilicoglu. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 239–243, 2003.

Marcelo Fiszman, Thomas C. Rindflesch, and Halil Kilicoglu. Abstraction summarization for managing the biomedical research literature. In *In Proceedings of the NAACL-HLT workshop on Computational Lexical Semantics*, pages 76–83, 2004.

Marcelo Fiszman, Dina Demner-Fushman, Halil Kilicoglu, and Thomas C. Rindflesch. Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of Biomedical Informatics*, 42(5):801–813, 2009.

Carol Friedman. Knowledge management and datamining in biomedicine. chapter Semantic Text Parsing for patient records, pages 423–448. Springer New York, 2005.

Carol Friedman and George Hripcsak. Natural language processing and its future in medicine. *Academic Medicine*, 74(8):890–893, August 1999.

Robert Gaizauskas, Patrick Herring, Michael Oakes, Michelline Beaulieu, Peter Willett, Helene Fowkes, and Anna Jonsson. Intelligent access to text: Integrating information extraction technology into text browsers. In *Proceedings of the Human Language Technology Conference (HLT)*, 2001.

Amit X. Garg, Neill K. J. Adhikari, Heather McDonald, Patricia Rosas-Arellano, P. J. Devereaux, Joseph Beyene, Justina Sam, and Brian Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *The Journal of The American Medical Association (JAMA)*, 293:1223–1238, March 2005.

Simon Gilbody. Evidence-based medicine. an improved format for journal clubs. *Psychiatric Bulletin*, 20:673–675, 1996.

Fiona Godlee. Getting evidence into practice. *BMJ*, 317:6, 1998.

Thomas Goetz and Claus-Wilhelm von der Lieth. PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic Acids Research*, 33:W774–W778, 2005.

Paul N. Gorman and Mark Helfand. Information seeking in primary care: How physicians choose which clinical questions to pursue and which to leave unanswered. *Medical Decision Making*, 15(2):113–119, 1995.

Andrew J. Graham and Sean C. Grondin. *Difficult Decisions in Thoracic Surgery*, chapter Evidence-Based Medicine: Levels of Evidence and Grades of Recommendation, pages 13–20. Springer London, 2007.

Michael L. Green and Tanya R. Ruff. Why do residents fail to answer their clinical questions? a qualitative study of barriers to practicing evidence-based medicine. *Academic Medicine: Journal of the Association of American Medical Colleges*, 80(2):176–182, February 2005.

Trisha Greenhalgh. Narrative based medicine in an evidence based world. *BMJ*, 318:323–325, 1999.

Trisha Greenhalgh. *How to read a paper: The Basics of Evidence-based Medicine*. Blackwell Publishing, 3 edition, 2006.

Udo Hahn and Michael Strube. Centering in-the-large computing referential discourse segments. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 104–111, 1997.

Udo Hahn, Martin Romacker, and Stefan Schulz. MEDSYNDIKATE — a natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics*, 67(1–3):63–74, 2002.

Sanda Harabagiu and Finley Lacatusu. Topic themes for multi-document summarization. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 202–209, 2005.

R. Brian Haynes, Nancy Wilczynski, K. Ann McKibbon, Cynthia J. Walker, and John C. Sinclair. Developing Optimal Search Strategies for Detecting Clinically Sound Studies in MEDLINE. *Journal of the American Medical Informatics Association (JAMIA)*, 1(6):447–458, 1994.

Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16. Association for Computational Linguistics, 1994.

Karin Friberg Heppin and Anni Jarvelin. Towards Improving Search Results for Medical Experts and Laypersons. In *Proceedings of CLEFeHealth*, 2012.

William R. Hersh, M. Katherine Crabtree, David H. Hickman, Lynetta Scherek, Charles P. Friedman, Patricia Tidmarsh, Craig Mosbaek, and Dale Kraemer. Factors associated with searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association (JAMIA)*, 9:283–293, 2002.

Arjen Hoogendam, Anton F. H. Stalenhoef, Pieter F. de Vries Robbé, and A. John P. M. Overbeke. Analysis of queries sent to pubmed at the point of care: observation of search behaviour in a medical teaching hospital. *BMC Medical Informatics and Decision Making*, 8:42, 2008.

Eduard Hovy and Chin-Yew Lin. Automatic Text Summarization and the SUMMARIST System. In *Proceedings of the TRIPSTER Text Program: Phase III*, pages 197 – 214, 1998.

Eduard Hovy and Chin-Yew Lin. *Advances in Automatic Text Summarization*, chapter Automated Text Summarisation in SUMMARIST, pages 81–94. MIT Press, 1999.

Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. Evaluation of PICO as a knowledge representation for clinical questions. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 359–363, 2006.

Dereck L. Hunt and K. Ann McKibbon. Locating and appraising systematic reviews. *Annals of Internal Medicine*, 126(7):532–538, 1997.

Lawrence Hunter and K. Bretonnel Cohen. Biomedical language processing: what's beyond pubmed. *Mollecular Cell*, 21:589–594, 2006.

Richard B. Ismach. Teaching evidence-based medicine to medical students. *Academic Emergency Medicine: Official Journal of the Society of Academic Emergency Medicine*, 11:e6–10, 2004.

Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, pages pp. 19–33, 1997.

David B. Johnson, Qinghua Zhu, John D. Dionisio, Victor Zhenyu Liu, and Wesley W. Chu. Modeling medical content for automated summarization. *Annual New York Academy of Sciences*, 980:247–258, 2002.

Desmond Jordan, Gregory Whalen, Blaine Bell, Kathleen McKeown, and Steven Feiner. *MEDINFO 2004*, chapter An Evaluation of Automatically Generated Briefings of Patient Status, pages 227–231. Amsterdam: IOS Press, 2004.

Mijail Kabadjov, Josef Steinberger, Ralf Steinberger, Massimo Poesio, and Bruno Pouliquen. Enhancing n-gram-based summary evaluation using information content and a taxonomy. In *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 662–666. Springer Berlin / Heidelberg, 2010.

Michael W. Kalsman and David A. Acosta. Use of the internet as a medical resource by rural physicians. *Journal of the American Board of Family Practice*, 13:349–352, 2000.

Sarvnaz Karimi, Justin Zobel, Stefan Pohl, and Falk Scholer. The challenge of high recall in biomedical systematic search. In *Proceedings of the third international workshop on Data and text mining in bioinformatics*, pages 89–92, 2009.

David L. Katz. *Clinical epidemiology & evidence-based medicine: fundamental principles of clinical reasoning and research*. SAGE, illustrated edition, 2001.

Liadh Kelly, Sebastian Dungs, Sascha Kriewel, Allan Hanbury, Lorraine Goeuriot, Gareth J. F. Jones, Georg Langs, and Henning Muller. Professional: Multilingual, Multimodal Professional Medical Search. In *ECIR 2014*, pages 754 – 758, 2014.

Halil Kilicoglu, Dina Demner-Fushman, Thomas C. Rindflesch, Nancy L. Wilczynski, and Brian R. Haynes. Towards automatic recognition of scientifically rigorous clinical research evidence. *Journal of the American Medical Informatics Association (JAMIA)*, 16(1):25–31, January 2009.

Su N. Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. Automatic classification of sentences to support Evidence Based Medicine. *BMC bioinformatics*, 12 Suppl 2, 2011.

Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, 1995.

Finley Lacatusu, Paul Parker, and Sanda Harabagiu. Lite-gistexter: Generating short summaries with minimal resources. In *Proceedings of the Document Understanding Conference*, pages 122–128, 2003.

J. Richard Landis and Gary G Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, March 1977.

S. Le Cessie and J. C. Van Houwelingen. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1):191–201, 1992.

Minsuk Lee, James Cimino, Hai Ran Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. Beyond information retrieval – medical question answering. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 469–473, 2006a.

Minsuk Lee, Weiqing Wang, and Hong yu. Exploring supervised and unsupervised methods to detect topics in biomedical text. *BMC Bioinformatics*, 7:140, 2006b.

Alessandro Lenci, Roberto Bartolini, Nicoletta Calzolari, Ana Agua, Stephan Busemann, Emmanuel Cartier, Karine Chevreau, and José Coch. Multilingual Summarization by Integrating Linguistic Resources in the MLIS-MUSI Project. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1464–1471, 2002.

Jurij Leskovec, Natasa Milic-Frayling, and Marko Grobelnik. Impact of linguistic analysis on the semantic graph coverage and learning of document extracts. In *Proceedings of AAAI*, pages 1069–1074, 2005.

Chin-Yew Lin. Training a selection function for extraction. In *Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM)*, pages 1–8, 1999.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of NAACL-HLT*, 2004.

Chin-Yew Lin and Eduard Hovy. Identifying topics by position. In *Proceedings of the Fifth conference on Applied Natural Language Processing*, pages 283–290, 1997.

Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 495–501, 2000.

Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, 2002.

Chin-Yew Lin and Eduard Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of NAACL-HLT*, pages 71–78, 2003.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of NAACL-HLT*, pages 463–470, 2006.

Jimmy J. Lin and Dina Demner-Fushman. The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 99–106, 2006.

Jimmy J. Lin and Dina Demner-Fushman. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

Donald A. Lindberg, Betsy L. Humphreys, and Alexa McCray. The unified medical language system. *Methods of Information in Medicine*, 32:281–291, 1993.

Marina Litvak and Mark Last. Graph-Based Keyword Extraction for Single-Document Summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, pages 17–24, 2008.

Annie Louis and Ani Nenkova. Automatic Summary Evaluation without Human Models. Technical report, 2008.

Annie Louis and Ani Nenkova. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42, 2011.

H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal*, 2:159–165, 1958.

Inderjeet Mani. *Automatic Summarization*. John Benjamins Publishing Company, 2001.

Inderjeet Mani and Eric Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of AAAI*, pages 622–628, 1997.

Inderjeet Mani and Mark T. Maybury. *Advances in Automatic Text Summarization*. MIT Press, 1999.

Inderjeet Mani, Mark T. Maybury (editors, and Mark Sanderson). *Book Review: Advances in Automatic Text Summarization edited by Inderjeet Mani and Mark T. Maybury*. MIT Press, 2000.

Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundhem. SUMMAC: A text summarisation evaluation. *Natural Language Engineering*, 8(1): 43–68, 2002.

Daniel Marcu. *Intelligent scalable text summarisation*, chapter From discourse structures to text summaries, pages 82–88. Proceedings of a Workshop Sponsored by the Association for Computational Linguistics, 1997.

Daniel Marcu. To build text summaries of high quality, nuclearity is not sufficient. In *Proceedings of AAAI*, pages 1–8, 1998.

Daniel Marcu. *Advances in Automatic Text Summarization*, chapter Discourse Trees are Good Indicators of Importance in Text, pages 123–136. MIT Press, 1999.

Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarisation*. Cambridge MA: MIT Press, 2000.

Mark T. Maybury. Generating summaries from event data. *Information Processing and Management*, 31(5):735–751, September 1995.

Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI Workshop on Learning for Text Summarization*, pages 41–48, 1998.

Alastair McColl, Helen Smith, Peter White, and Jenny Field. General practitioner's perceptions of the route to evidence based medicine: a questionnaire survey. *BMJ*, 316:361–365, 1998.

Clement J. McDonald. Protocol-based computer reminders, the quality of care and the non-perfectability of man. *The New England Journal of Medicine*, 295(24):1351–1355, 1976.

Bridget T. McInnes, Ted Pedersen, and Serguei V. S. Pakhomov. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 431–435, 2009.

Kathleen R. McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, 1995.

Kathleen R. McKeown, Jacques Robin, and Karen Kukich. Generating concise natural language summaries. *Information Processing and Management*, 31(5):703–733, September 1995.

Kathleen R. McKeown, Desmond A. Jordan, and Vasileios Hatzivassiloglou. Generating patient-specific summaries of online literature. Technical report, AAAI Technical Report SS-98-06, 1998.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the second international conference on Human Language Technology Research*, pages 280–285, 2002.

Gabor Melli, Zhongmin Shi, Yang Wang, Yudong Liu, Anoop Sarkar, and Fred Popowich. The SFU Question Answering Summary Handler. In *Proceedings of the Document Understanding Conference (DUC)*, pages 103–110, 2005.

Frank Meng, Ricky K. Taira, Alex A. T. Bui, Hooshang Kangarloo, and Bernard M. Churchill. Automatic generation of repeated patient information for tailoring clinical notes. *International Journal of Medical Informatics*, 74:663–673, 2005.

Rada Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 2004.

Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

Seiji Miike, Etsuo Itoh, Kenji Ono, and Kazuo Sumita. A full-text retrieval system with a dynamic abstract generation function. In *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 152–161, 1994.

Randolph A. Miller, Harry E. Pople, and Jack D. Myers. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *The New England Journal of Medicine*, 307(8):468–478, 1982.

Jean-Luc Minel, Sylvaine Nugier, and Gérald Prat. How to Appreciate the Quality of Automatic Text Summarization? Examples Of FAN And MLUCE Protocols And Their Results On SERAPHIN. In *Proceedings of the Workshop On Intelligent Scalable Text Summarization, ACL-97*, pages 25–30, 1997.

Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. Abstracting of legal cases: the salomon experience. In *ICAIL '97: Proceedings of the 6th international conference on Artificial intelligence and law*, pages 114–122, 1997.

Diego Mollá. A Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Association Workshop*, volume 8, 2010.

Diego Mollá and Maria Elena Santiago-Martinez. Development of a Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, December 2011.

Diego Mollá and Abeed Sarker. Automatic Grading of Evidence: the 2011 ALTA Shared Task. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 4–8, 2011.

Diego Mollá and José Luis Vicedo. Question Answering in Restricted Domains: An Overview. *Computational Linguistics*, 33:41–61, 2007.

Victor M. Montori and Gordon H. Guyatt. Progress in evidence-based medicine. *The Journal of The American Medical Association (JAMA)*, 300:1814–1816, 2008.

Victor M. Montori, Nancy L. Wilczynski, Douglas Morgan, and R. Brian Haynes. Optimal search strategies for retrieving systematic reviews from medline: analytical survey. *BMJ*, 330(7482): 68–73, 2005.

Andre H. Morris, George M. Kasper, and Dennis A. Adams. The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research*, 3(1):17–35, 1992.

Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of NAACL-HLT*, 2004.

Ani Nenkova and Rebecca Passonneau. The impact of frequency on summarization. MSR-TR, Microsoft Research, Redmond, Washington, 2005.

Yun Niu. *Analysis of Semantic Classes: Toward Non-Factoid Question Answering*. PhD thesis, University of Toronto, 2007.

Yun Niu and Graeme Hirst. Analysis of semantic classes in medical text for question answering. In *Proceedings of the ACL-2004 workshop Question Answering in Restricted Domains, Barcelona, Spain*, 2004.

Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. Answering Clinical Questions with Role Identification. In *Proceedings of the ACL-2003 workshop Natural Language Processing in Biomedicine*, 2003.

Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. Analysis of polarity information in medical text. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 570–574, 2005.

Yun Niu, Xiaodan Zhu, and Graeme Hirst. Using outcome polarity in sentence extraction for medical question-answering. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 599–603, 2006.

Miles Osborne. Using maximum entropy for sentence extraction. In *Proceedings of the ACL'02 Workshop on Automatic Summarization*, pages 1–8, 2002.

Norman Papernick and Alexander G. Hauptmann. Summarization of broadcast news video through link analysis of named entities. In *AAAI workshop on Link Analysis*, 2005.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Aytomatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, number 311–318, 2002.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, pages 185–208, Cambridge, MA, 1998. MIT Press.

Laura Plaza, Alberto Diaz, and Pablo Gervas. A semantic graph-based approach to biomedical summarisation. *Artificial Intelligence in Medicine*, 53:1–14, 2011a.

Laura Plaza, Antonio Jimeno-Yepes, Alberto Diaz, and Alan Aronson. Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC Bioinformatics*, 12(1):355–368, 2011b.

Maksim V. Plikus, Zina Zhang, and Cheng-Ming Chuong. Pubfocus: semantic medline/pubmed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics*, 7(1):424–439, 2006.

Stefan Pohl, Justin Zobel, and Alistair Moffat. Extended boolean retrieval for systematic biomedical reviews. In *Proceedings of the thirty-third Australasian Computer Science Conference*, 2010.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. A rule-based approach to discourse parsing. In *Proceedings of the fifth SIGdial workshop on Discourse and Dialogue*, pages 108–117, 2004.

Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

John R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

Dragomir R. Radev and Kathleen R. Mckeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):470–500, 1998.

Dragomir R. Radev, Hongyang Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, pages 21–30, 2000a.

Dragomir R. Radev, John M. Prager, and Valerie Samn. Ranking suspected answers to natural language questions using predictive annotation. In *Proceedings of the sixth conference on Applied Natural Language Processing*, pages 150–157, 2000b.

Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408, 2002.

Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938, 2004.

Lawrence H. Reeve, Hyoil Han, and Ari D. Brooks. Biochain: Using lexical chaining methods for biomedical text summarization. In *Proceedings of the 21st annual ACM symposium on applied computing, bioinformatics track*, pages 180–184, 2006a.

Lawrence H Reeve, Hyoil Han, Saya V. Nagori, Jonathan C. Yang, Tamara A. Schwimmer, and Ari D. Brooks. Concept frequency distribution in biomedical text summarization. In *Proceedings of the ACM 15th conference on information and knowledge management (CIKM'06)*, pages 604–611, 2006b.

Lawrence H. Reeve, Hyoil Han, and Ari D. Brooks. The use of domain-specific concepts in biomedical text summarization. *Information Processing and Management*, 43:1765–1776, 2007.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial Intelligence (IJCAI'95)*, volume 1, 1995.

Scott W. Richardson, Mark C. Wilson, Jim Nishikawa, and Robert S. Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–A13, 1995.

Thomas C. Rindflesch, Marcelo Fiszman, and Bisharah Libbus. Chapter 14 semantic interpretation for the biomedical research literature, 2005.

David L. Sackett, Brian R. Haynes, Gordon H. Guyatt, and Peter Tugwell. *Clinical epidemiology: A basic science for clinical medicine*. Little Brown & Co. Inc., 2 edition, 1991.

David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, Brian R. Haynes, and W. Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996. URL http://www.bmj.com.

David L. Sackett, Sharon E. Straus, and W. Scott Richardson. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, 2 edition, 2000.

Kenji Sagae and Jun'ichi Tsujii. Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 1044–1050, 2007.

Naomi Sager, Margaret Laman, Christine Bucknall, Ngo Nhan, and Leo J. Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association (JAMIA)*, 1:142–160, 1994.

Horacio Saggion and Guy Lapalme. Generating indicative-informative summaries with sumum. *Computational Linguistics*, 28(4):497–526, 2002.

Gerard Salton and Michael J. McGill. *Introduction to modern information retrieval*. McGraw Hill, New York, NY, USA, 1983.

Tom Sanders, Stephen Harrison, and Kath Checkland. Evidence-based medicine and patient choice: the case of heart failure care. *Journal of Health Services Research and Policy*, 13: 103–108, 2008.

Kamal Sarkar. Using domain knowledge for text summarization in medical domain. *International Journal of Recent Trends in Engineering (IJRTE)*, 1(1):200–205, 2009.

Abeed Sarker, Diego Mollá, and Cécile Paris. Outcome Polarity Identification of Medical Papers. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 105–114, December 2011.

Abeed Sarker, Diego Mollá, and Cécile Paris. An approach for query-focused text summarisation for evidence based medicine. In Niels Peek, Roque Martin Morales, and Mor Peleg, editors, *Artificial Intelligence in Medicine*, volume 7885 of *Lecture Notes in Computer Science*, pages 295–304. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-38325-0. doi: 10.1007/978-3-642-38326-7_41. URL http://dx.doi.org/10.1007/978-3-642-38326-7_41.

Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the ACL*, 2009.

Frank Schilder and Ravikumar Kondadadi. Fastsum: Fast and accurate query-based multi-document summarization. In *Proceedings of ACL-HLT, Short Papers*, pages 205–208, 2008.

Ralf W. Scholosser, Rajinder Koul, and John Costello. Asking well-built questions for evidence-based practice in augmentative and alternative communication. *Journal of Communication Disorders*, 40:225–238, 2006.

Rolf Schwitter. Creating and Querying Formal Ontologies via Controlled Natural Language. *Applied Artificial Intelligence*, 24(1-2):149–174, 2010.

Charlotte Seckman, Dina Demner-Fushman, Cheryl Fisher, S. C. Hauser, and George Thoma. InfoBot: A Prototype System to Support Evidence-based Practice. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 151–155, 2008.

Sanchaya Selvaraj, Yeshwant Kumar, Elakiya, Prarthana Saraswathi, Balaji, Nagamani, and SuraPaneni Krishna Mohan. Evidence-based medicine - a new approach to teach medicine: a basic review for beginners. *Biology and Medicine*, 2(1):1–5, 2010.

Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M. Kashani, Anoop Sarkar, and Fred Popowich. Question answering summarization of multiple biomedical documents. In *Proceedings of the 20th Canadian Conference on Aritificial Intelligence (CanAI '07)*, 2007.

Kaveh G. Shonjania and Lisa A. Bero. Taking advantage of the explosion of systematic reviews: An efficient medline search strategy. *Effective Clinical Practice*, 4(4):157–162, 2001.

Edward H. Shortliffe. Computer Programs to Support Clinical Decision Making. *The Journal of the American Medical Association*, 258(1):61–66, 1990.

Edward H. Shortliffe, Bruce G. Buchanan, and Edward A. Feigenbaum. Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. In *Proceedings of the IEEE*, volume 67. 1207–1223, 1979.

C. Silagy, D. Weller, P. Middleton, and J. Doust. General practioners' use of evidence databases. *Medical Journal of Australia*, 170:393, 1999.

David C. Slawson and Allen F. Shaughnessy. Teaching Evidence-Based Medicine: Should We Be Teaching Information Management Instead? *Academic Medicine*, 80(7):685–689, 2005.

Richard Smith. What clinical information do doctors need. *BMJ*, 313:1062–1068, 1996.

Karen Sparck Jones. Automatic summarizing: factors and directions. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in automatic text summarization*, chapter 1, pages 1 – 12. The MIT Press, 1999.

Karen Sparck Jones. Factorial summary evaluation. Invited talk at the Workshop on Text Summarisation, ACM SIGIR 2001 Conference, 2001.

Karen Sparck Jones. Automatic summarising: The state of the art. *Information Processing and Management*, 43:1449 – 1481, 2007.

Hanna Suominen, Sampo Pyysalo, Marketta Hissa, Filip Ginter, Shuhua Liu, Dorina Marghescu, Tapio Pahikkala, Barbro Back, Helena Karsten, and Tapio Salakoski. *Handbook of Research on Text and Web Mining Technologies*, chapter Performance Evaluation Measures for Text Mining, pages 724 – 747. IGI Global, 2008.

Hanna Suominen, Sanna Salantera, Sumitra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In Pamela Forner, Henning Muller, Roberto Predes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 212–231. 2013.

Charles Sutton and Andrew McCallum. *Introduction to Statistical Relational Learning*, chapter Introduction to Conditional Random Fields for Relational Learning, pages 93–128. MIT Press, 2007.

Krysta M. Svore, Lucy Vanderwende, and Christopher J. C. Burges. Enhancing single-document summarization by combining ranknet and third-party sources. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 448–457, 2007.

Thanh Tang, David Hawking, Ramesh Sankaranarayana, Kathleen Griffiths, and Nick Craswell. Quality-Oriented Search for Depression Portals. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, chapter 60, pages 637–644. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2009.

Ross J. Taylor, Brian R. McAvoy, and Tom O'Dowd. *General Practice Medicine: an illustrated colour text*. Elsevier Health Sciences, 2003.

Rafael M. Terol, Patricio Martínez-Barco, and Manuel Palomar. Applying nlp techniques and biomedical resources to medical questions in qa performance. In *Proceedings of the Mexican International Conference on Artificial Intelligence*, pages 996–1006, 2006.

Rafael M. Terol, Patricio Martínez-Barco, and Manuel Palomar. A knowledge based method for the medical question answering problem. *Computer Methods and Programs in Biomedicine*, 37(10):1511–1521, 2007.

Simon Teufel. Task-based evaluation of summary quality: Describing relationships between scientific papers. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*, pages 12–21, 2001.

Simon Teufel and Marc Moens. Sentence extraction as a classification task. In *Proceedings of the ACL-97*, pages 58–65, 1997.

Gian Lorenzo Thione, Martin Van den Berg, Livia Polanyi, and Chris Culy. Hybrid text summarization: Combining external relevance measures with structural analysis. In Stan Szpakowicz Marie-Francine Moens, editor, *Proceedings of the Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 51–55, 2004.

Andreea Tutos and Diego Mollá. A study on the use of search engines for answering clinical questions. In *Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management*, pages 61–68, 2010.

Ozlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association (JAMIA)*, 16:109–115, 2009.

Hans van Halteren and Simone Teufel. Examining the concensus between human summaries: initial experiments. In *Proceedings of the NAACL-HLT Workshop on Text summarization*, volume 5, pages 57–64, 2003.

Mathias Verbeke, Vincent Van Asch, Roser Morante, Paolo Frasconi, Walter Daelemans, and Luc De Raedt. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings EMNLP-CoNLL*, pages 579–589, 2012.

Anita A. H. Verhoeven, Edzard J. Boerma, and Betty Meyboom de Jong. Which literature retrieval method is most effective for gps? *Family Practice*, 17(1):30–35, 2000.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9 (Suppl 11)(S9), 2008.

Tielman T. Van Vleck and Noemie Elhadad. Corpus-Based Problem Selection for EHR Note Summarization. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 817–821, 2010.

Tielman T. Van Vleck, Daniel M. Stein, Peter D. Stetson, and Stephen B. Johnson. Assessing data relevance for automated generation of clinical summary. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 761–765, 2007.

Wang Weiming, Dawei Hu, Min Feng, and Liu Wenyin. Automatic clinical question answering based on UMLS relations. In *Proceedings of the third International Conference on Semantics, Knowledge and Grid*, 2007.

Barry D. Weiss. SORT: Strength of Recommendation Taxonomy. *Family Medicine*, 36(2): 141–143, February 2004.

Suzanne West, Valerie King, Timothy S. Carey, Kathleen N. Lohr, Nikki McKoy, Sonya F. Sutton, and Linda Lux. Systems to rate the strength of scientific evidence. http://www.arhq.gov/clinic/epcsums/strengthsum.htm, April 2002. URL http://www.arhq.gov/clinic/epcsums/strengthsum.htm.

Edward E. Westberg and Randolph A Miller. The basis for using internet to support the information needs of primary care. *Journal of the American Medical Informatics Association (JAMIA)*, 6:6–25, 1999.

Michael White and Claire Cardie. Selecting sentences for multidocument summaries using randomized local search. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, volume 4, pages 9–18, 2002.

P. A. Williams and S. P. Maj. Is the internet an integral part of general practice in australia? *Studies in health technology and informatics*, 84(Pt. 1):394–398, 2001.

John W. Williamson, Pearl S. German, Robin Weiss, Elizabeth A. Skinner, and Frederick Bowes, III. Health science information management and continuing education of physicians. *Annals of Internal Medicine*, 110(2):151 – 160, 1989.

Paul Wilson, Julian Droogan, Julie Glanville, Ian Watt, and G. Hardman. Access to the evidence base from general practice: a survey of general practice staff in Northern and Yorkshire Region. *Quality in Health Care*, 10(2):83–86, 2001.

Steven M. Wilson. Impact of internet on primary care staff in glasgow. *Journal of Medical Internet Research*, 1(2):E7, 1999.

T. Elizabeth Workman, Marcelo Fiszman, and John F. Hurdle. Text summarisation as a decision support aid. *BMC Medical Informatics and Decision Making*, 12:41–53, 2012.

Hua Xu, Shane P. Stenner, Son Doan, Kevin B. Johnson, Lemuel R. Waltman, and Joshua C. Denny. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association (JAMIA)*, 17:19–24, 2010.

Wen-Tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. Multi-document summarization by maximizing informative content-words. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1776–1782, 2007.

Jane M. Young and Jeanette E. Ward. Evidence-based medicine in general practice: Beliefs and barriers among australian gps. *Journal of Evaluation in Clinical Practice*, 7(2):201–210, 2001.

Hong Yu and Yong-gang Cao. Automatically extracting information needs from ad hoc clinical questions. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pages 96–100, 2008.

Hong Yu and David Kaufman. A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. *Pacific Symposium on Biocomputing*, 12:328–339, 2007.

Hong Yu and Carl Sable. Being Erlan Shen: Identifying Answerable Questions. In *Proceedings of the IJCAI'05 Workshop on Knowledge and Reasoning for Answering Questions (KRAQ'05)*, pages 6–14, 2005.

Hong Yu, Carl Sable, and Hai Zhu. Classifying medical questions based on an evidence taxonomy. In *Proceedings of the AAAI Workshop Question Answering in Restricted Domains*, pages 27–35, 2005.

Hong Yu, Minsuk Lee, David Kaufman, John Ely, Jerome A. Osheroff, George Hripcsak, and James Cimino. Development, implementation and, a cognitive evaluation of a definitional question answering system for physicians. *Journal of Biomedical Informatics*, 40:236–251, 2007a.

Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13:25–49, 2007b.

Qing T. Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N. Myurphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6:30–38, 2006.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of NAACL-HLT*, pages 447–454, 2006.

Pierre Zweigenbaum. Question answering in biomedicine. In *Proceedings of the EACL 2003 Workshop on Natural Language Processing for Question Answering*, 2003.

Pierre Zweigenbaum. Knowledge and reasoning for medical question-answering. In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions, ACL-IJCNLP 2009*, pages 1–2, 2009.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375, September 2007.