Appraising UMLS Coverage for Summarizing Medical Evidence

Elaheh ShafieiBavani Mohammad Ebrahimi Raymond Wong

Fang Chen

University of New South Wales, Australia
Data61, CSIRO, Australia
{elahehs,mohammade,wong,fang}@cse.unsw.edu.au

Technical Report
UNSW-CSE-TR-201611
July 2016

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

When making clinical decisions, practitioners need to rely on the most relevant evidence available. However, accessing a vast body of medical evidence and confronting the issue of information overload, can be challenging and time consuming. Automatic text summarization has been known as a natural language processing technique to address this issue. While most top-performing summarizers remain largely extractive (i.e. extract a group of sentences and concatenate them.), this paper proposes an abstractive query-focused summarization framework for evidence-based medicine (EBM). Given a clinical query and a set of relevant medical evidence, our aim is to generate a fluent, well-organized, and compact summary that answers the query. The quality of biomedical summaries is also enhanced by appraising the applicability of both general-purpose (WordNet), and domain-specific (UMLS) knowledge sources for concept discrimination.

We first perform iterative random walks, over the graph representation of both WordNet and UMLS, to capture sentence-to-query and sentence-to-sentence semantic similarities. We then construct a similarity graph with less query-relevant sentences filtered out, and relevant sentences are clustered. Finally, a word graph is constructed for each cluster, and the most abstractive summary sentences are obtained by re-ranking k-shortest paths. Analysis via ROUGE metrics shows that using WordNet as a general-purpose lexicon helps to capture the concepts not covered by the UMLS Metathesaurus, and hence significantly increases the summarization performance. The effectiveness of our proposed framework is demonstrated by conducting a set of experiments over a specialized EBM corpus - which has been gathered and annotated for the purpose of biomedical text summarization.

1 Introduction

Over the past two decades, clinical guidelines urged practitioners to move towards evidence-based medicine, which is formally defined as *conscientious and judicious use of current best evidence in making decisions about the care of individual patients* [50]. Evidence-based medical practice heavily relies on research evidence, rather than intuition, unsystematic clinical experience, or pathologic rationale [19]. So, the main aim is to find and evaluate current medical evidence, and make clinical decisions based on the best available evidence. Systematic reviews, meta-analyses of all studies associated with a topic, or high-quality randomized controlled trials, are widely known as the best sources of evidence [34].

However, searching through and evaluating primary medical literature is extremely time consuming [13, 51, 24]. For more clarity, a query on PubMed[1] [32], returns a large set of relevant documents, and not summaries or answers to the queries. Even targeted searches on PubMed tend to return a large volume of results. Hence, the explosive growth of content of medical evidence requires development of techniques to present information to physicians and researchers in an effective way. Automatic text summarization has been introduced as a natural language processing technique to address this problem [26, 49, 17]. Well-generated summaries can efficiently reduce diagnosis time, and help physicians to identify treatment options [7]. Moreover, automatic summaries have been shown to improve indexing and categorizing biomedical literature when used as substitutes for the articles abstracts [18, 27].

Even though the problem of information overload and the advantages of summarization are critical in the biomedical domain, the majority of summarizers are designed to be general-purpose, and do not take into account the particular properties of specific domains like biomedical. They usually work with a simple representation of the summary comprising of information that can be directly extracted from the document itself, such as terms, phrases or sentences [14, 36]. However, recent studies (e.g. [17]) have demonstrated the benefits of summarization based on richer representations that make use of domain-specific knowledge sources. These approaches represent the documents using concepts instead of words, and may also be enriched by using semantic associations among concepts (e.g. synonymy, hypernymy, homonymy or co-occurrence) [47]. While a query is asked in the field of biomedicine, one of their main challenges is to understand the underlying semantic relatedness of the query and document sentences, and consequently extract the most non-redundant, query-relevant parts from the documents [37].

Documents in biomedicine are very different from documents in other fields, and include very different document types (e.g. patient records, web documents, scientific papers and even e-mailed reports). So, particular characteristics of the domain and the type of documents are apparently needed to be considered [47]. In addition, medical language, despite being highly specialized, is also highly interpretive, and it is constantly expanding [47]. It seems reasonable that these peculiarities should be exploited by the summarization systems. To this end, promising domain-specific NLP techniques have been efficiently employed to release the Unified Medical Language System (UMLS[2]) [5]. UMLS covers a wide range of biomedical concepts, semantic types, and both hierarchical and non-hierarchical relationships among the concepts.

[1] A biomedical database produced by the U.S. National Library of Medicine, and contains over 24 million articles (available at http://www.ncbi.nlm.nih.gov/pubmed).

[2] A repository of biomedical vocabularies, were developed by the U.S. National Library of Medicine (available at http://www.nlm.nih.gov/research/umls/).

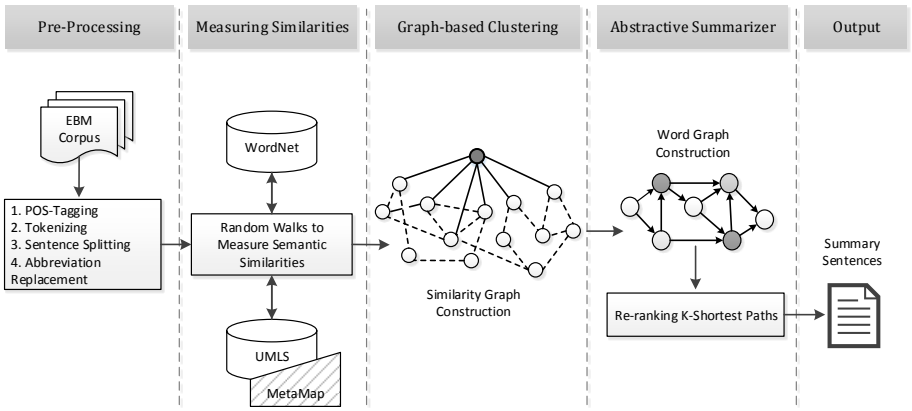


Figure 1.1: An overview of the proposed framework

More specifically, it was developed to overcome the main issue in this area, which is the absence of a standard format for distributing terminologies [5]. UMLS is a very rich source of information in medical and biological domain. Therefore, most existing biomedical summarizers utilize UMLS as a large lexical and semantic medical ontology. However, UMLS does not provide a full coverage of non-medical concepts, terms, and relations included in general-purpose thesauri such as WordNet[3] [23, 8, 25, 41]. Moreover, utilizing WordNet to complement UMLS coverage is challenging due to their different structure, nature, terms and size.

This challenge has motivated us to propose an efficient summarization framework to tackle the aforementioned issues by keeping an eye on the biomedical peculiarities. Given a clinical inquiry and a set of relevant medical evidence, our abstractive summarization framework aims to generate a fluent, well-organized, and compact summary that answers the query. To this end, we provide a deeper analysis of the biomedical text to generate new sentences that convey the gist of the source content. The quality of biomedical summaries is also enhanced by appraising the applicability of both general-purpose (WordNet), and domain-specific (UMLS) knowledge sources for concept discrimination. In details, our framework comprises different components: performing iterative random walks on WordNet and UMLS to capture the underlying sentence-to-query and sentence-to-sentence semantic similarities; ranking sentences based on their similarity scores; filtering them considering their relationship to the clinical query; clustering them by their relevance to each other; generating abstractive summarization, and removing redundancies through a word graph representation; and finally re-ranking the newly generated sentences based upon their importance and syntactic structure. Figure 1.1 provides an overview of the proposed framework.

The rest of the paper is organized as follows. Section 2 summarizes the background. Utilized data is discussed in Section 3. Preprocessing step is explained in Section 4. We demonstrate the proposed approach in Section 5. Section 6 reports the evaluation metrics and the performed experiments. Finally, Section 7 concludes the paper.

[3]http://wordnet.princeton.edu

2 Background

2.1 Text Summarization

Text summarization is the process of automatically creating a compressed version of a given text. Content reduction can be addressed by selection and/or by generalization of what is important in the source [28]. Consequently, two common categories are defined in the text summarization literature:

- Extractive summarization: This category usually takes a set of documents as input and selects the salient sentences for inclusion in the final summary. Therefore, the summaries are essentially composed of material that is explicit in the source.
- Abstractive summarization: This category generates summaries in which the information from the source has been paraphrased. Human summaries are also typically abstracts.

A summary can either be query-focused (biased to a user query), or generic (conveying the document gist). In traditional extractive query-focused summarization systems, lexical similarity measures were used to select content that are similar to the question. Such approaches also have to ensure that redundant information is minimized. Some recent researches have addressed query-focused text summarization from the perspective of question answering [67, 65], and some others have modeled summarization as sentence classification problems [12, 44, 43, 9]. A machine learning classifier trained on a small dataset is employed in another study [12] to select the summary sentences. Another summarization system [9] utilizes category of an input question to generate paragraph level summaries. They suggest that the generated summary should be customized with respect to the type of the question. In biomedical summarization, polarity information of sentences is utilized by [44, 43] to summarize medical abstracts. They believe that polarized sentences could favorably conclude statements within the abstracts.

More advanced summarization techniques such as LexRank [14] incorporate graph-based methods. LexRank assumes a fully connected and undirected graph for the set of documents to be summarized. Each node corresponds to a sentence represented by its TF-IDF vector, and the edges are labeled with the cosine similarity between the sentences. Only the edges that connect sentences with a similarity above a predefined threshold are drawn in the graph. The sentences represented by the most highly connected nodes are selected for the summary. Recent extractive summarization approaches have also attempted more targeted tasks. They automatically assessed the risk of bias for clinical trials [34], and extracted specific study characteristics from trial abstracts [60].

So, the majority of proposed systems on general-purpose and domain-specific (e.g. biomedical) text summarization are extractive in nature. This is mainly due to the difficulties entailed by the abstraction process. This process usually involves identifying the most prevalent concepts in the source, the appropriate semantic representation of them, and rewriting of the summary through natural language generation techniques. In this paper, one of our main contributions is to provide an abstractive summary of a set of biomedical research evidence with respect to a clinical query.

2.2 Evidence-based Medical Summarization

Among the researches performed in the area of text summarization, many studies have also explored the obstacles associated with evidence-based medicine practice in the absence of pre-existing systematic reviews (e.g. [13, 11]). When primary care physicians seek answers to clinical problems, the time required to search, evaluate, and synthesize evidence has been known as a major obstacle [52]. Literature review and analysis may take a long time (e.g. it takes more than 30 minutes on average for a practitioner to find and extract evidence [22]). Due to the difficulties such as time needs, practitioners often do not pursue evidence-based answers to clinical problems. Besides, clinical resource developers need to focus on providing resources that answer questions likely to occur in practice with emphasis on treatment and bottom-line advice.

Numerous IR approaches have already been proposed to address the search-related needs of practitioners [20]. They incorporate lexical and semantic information derived from domain-specific resources and ontologies. However, post-retrieval techniques (e.g. [52]) to perform query-oriented summarization are still scarce. The possible reasons are as follows: (i) the complicated nature of the biomedical text that arises various difficulties in progress [3]; (ii) the large volume of available published medical literature; (iii) the limited amount of suitable annotated data for the complex task of summarization or question answering [52].

Due to the lack of incorporation of domain-specific information, domain-independent summarizers generally underperform when employed on biomedical texts. To solve this issue, UMLS came to play, and has proved to be a useful knowledge source for summarization in biomedical domain [49, 17]. However, a decline is found in the performance of summarizers which only utilize UMLS as their source of knowledge. The reason is that UMLS is less likely to cover all concepts included in the source text [47]. To compensate this deficiency, a question-oriented extractive system for biomedical multi-document summarization (i.e. [57]), utilized WordNet as a general-purpose lexicon to capture the concepts not covered by UMLS. For this intention, they constructed a graph containing ontological concepts (general ones from WordNet, and specific ones from UMLS), name entities, and noun phrases. The edges in their graph represent semantic relationships between concepts, but nothing is said about the considered specific relationships. Our work differs in intent, and explores the utility of the graph representation of both domain-independent (WordNet) and domain-specific (UMLS) lexicons for incorporating underlying textual semantic similarities. Next, we discuss these resources, their distinctions, and the employed EBM corpus.

3 Data

In our work, we have utilized two knowledge sources of UMLS and WordNet for concept discrimination. We have also employed the data provided in an EBM corpus to develop, test, and evaluate our summarization framework.

The Unified Medical Language System (UMLS) UMLS [5] is a database of biomedical vocabularies developed by the U.S. National Library of Medicine. In this work, we have utilized version 2015AB of the UMLS Metathesaurus that contains more than 3.25 million concepts, and nearly 13 million unique concept names from over 190 source vocabularies. The three major components of UMLS are the Metathesaurus, Semantic Network and SPECIALIST Lexicon. This work focuses on the Metathesaurus

which semi-automatically integrates information about biomedical and health-related concepts from various biomedical and clinical sources. UMLS uses 12 different types of hierarchical and non-hierarchical relations between concepts. For instance, the hierarchical relations consist of the *parent/child* and *broader/narrower* (BR/NR) relations. To map biomedical text to concepts in the UMLS Metathesaurus, MetaMap[1] program is usually applied [2]. MetaMap employs a knowledge-intensive approach that uses the SPECIALIST Lexicon in combination with lexical and syntactic analysis to identify noun phrases in text. We have employed version 2016 of MetaMap in our framework. Matches between noun phrases and Metathesaurus concepts are computed by generating lexical variations and allowing partial matches between the phrase and the concept. The possible UMLS concepts are assigned scores based on the closeness of the match between the input noun phrase and the target concept. The highest scoring concepts and their semantic types are gradually returned.

WordNet WordNet is a large general-purpose lexical database of English, which is often used in word sense discrimination. Words are grouped into sets of synonyms called synsets, each of which expressing a distinct concept. We have used WordNet 3.0 [15] repository for the current study, that includes a total of 155,287 words organized in 117,659 concepts, which are linked by semantic and lexical relations.

UMLS VS. WordNet Although WordNet includes a certain number of medical terms, in the area of biomedical, UMLS is used extensively for medical text mining and retrieval. A study performed by [5] showed that the concept overlap between WordNet and UMLS varies from 48% to 97%. This is because UMLS records the variability of the lexical forms encountered in the source vocabularies, while WordNet only records the canonical forms. WordNet and UMLS are also different in their graph structures. Therefore, there exists a huge discrepancy in granularity between WordNet and UMLS [33]. For example, as shown in Figure 3.1, *malignant_tumor.n.01* is the parent of *cancer.n.01* in WordNet, but "malignant tumor" and "cancer" locate in the same concept *C0006826* (malignant neoplasms) in UMLS. In this example, WordNet has a finer granularity. However, UMLS possesses a finer granularity in some other cases.

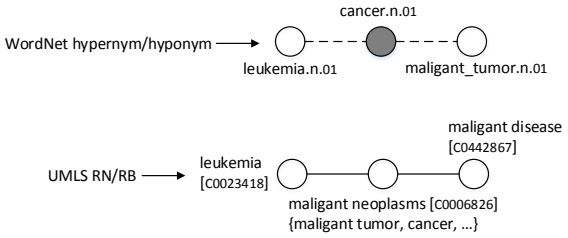


Figure 3.1: An Example of difference between WordNet and UMLS

While UMLS is a very rich source of information on medical and biological terms and concepts, it does not provide full coverage of non-medical concepts, terms and relations [23, 8, 25, 41]. In this paper, we have utilized WordNet to represent the layman knowledge, and UMLS to represent the professional knowledge. Our goal is to

[1]Developed by the U.S. National Library of Medicine (available at https://metamap.nlm.nih.gov)

capture sentence-to-query and sentence-to-sentence semantic similarities, and bridge the knowledge and language gaps in biomedical summarizers.

EBM Corpus To the best of our knowledge, the corpus released by [39] is the only available corpus[2] for the task of evidence-based medicine text summarization. This corpus is sourced from the Clinical Inquiries section of the Journal of Family Practice[3]. Each article in this section of the journal (issued monthly) addresses a clinical question, and provides a systematic analysis of the best available medical evidence in response to the posed clinical query [38]. For each question, this corpus contains the following information:

- The URL of the clinical inquiry: An address, from which the information has been sourced.
- The question: For example, "*What is the evaluation and treatment strategy for Raynaud's phenomenon?*".
- The bottom-line evidence-based answer: The answer may contain several parts, since a question may be answered according to distinct pieces of evidence. For each part, the corpus includes a short description of the answer, the Strength of Recommendation (SOR) grade of the evidence related to the answer, and a short description that explains the reasoning behind allocating such a SOR grade.
- The answer justifications: For each of the parts of the evidence-based answer, there is one or more justifications describing the actual findings reported in the research papers supporting the answer.
- The references: Each answer justification includes one or more references to the source research paper. Each reference includes the PubMed ID and full abstract information as encoded in PubMed, if available.

This corpus consists of 456 clinical queries, with 1396 bottom-line, multi-document summaries (i.e. evidence-based answers). The total number of single-document evidence-based summaries is 3036, which are generated from 2908 unique articles. The corpus also contains XML versions of these articles, obtained from PubMed. We have utilized this corpus to develop and test our query-focused multi-document summarization framework. The bottom-line answers are used as the reference (gold) summaries. The question and all the abstracts associated with the bottom-line summary are also considered as the source texts. Table 3.1 lists the properties of this corpus, and Table 3.2 provides an example of query-focused multi-document summarization over this corpus.

total #clinical queries	456
#bottom-line multi-document summaries	1396
#single-document evidence summaries	3036
total #unique articles	2908

Table 3.1: Information about the EBM Corpus

[2] Available at: http://sourceforge.net/projects/ebmsumcorpus
[3] http://www.jfponline.com/articles/clinical-inquiries.html

Question: How should we manage a patient with a positive PPD and prior BCG vaccination?
Bottom-line answer (multi-document summary): A recently developed alternative is the interferon-gamma assay (QuantiFERON-TB Gold test), which may be used in place of, or in addition to, the PPD skin test for patients who are known to have received a BCG vaccine. [*PubMed IDs: 15059788, 16539718*]
Source text 1 [PMID: 15059788]: The tuberculin skin test for immunologic diagnosis of Mycobacterium tuberculosis infection has many limitations, including being confounded by bacillus Calmette-Gurin (BCG) vaccination or exposure to nontuberculous mycobacteria. M. tuberculosis-specific antigens that are absent from BCG and most nontuberculous mycobacteria have been identified. We examined the use of two of these antigens, CFP-10 and ESAT-6, in a whole blood IFN-gamma assay as a diagnostic test for tuberculosis in BCG-vaccinated individuals. Because of the lack of an accurate standard with which to compare new tests for M. tuberculosis infection, specificity of the whole blood IFN-gamma assay was estimated on the basis of data from people with no identified risk for M. tuberculosis exposure (216 BCG-vaccinated Japanese adults) and sensitivity was estimated on the basis of data from 118 patients with culture-confirmed M. tuberculosis infection who had received less than 1 week of treatment. Using a combination of CFP-10 and ESAT-6 responses, the specificity of the test for the low-risk group was 98.1% and the sensitivity for patients with M. tuberculosis infection was 89.0%. The results demonstrate that the whole blood IFN-gamma assay using CFP-10 and ESAT-6 was highly specific and sensitive for M. tuberculosis infection and was unaffected by BCG vaccination status.
Source text 2 [PMID: 16539718]: The whole-blood interferon-gamma release assay (IGRA) is recommended in some settings as an alternative to the tuberculin skin test (TST). Outcomes from field implementation of the IGRA for routine tuberculosis (TB) testing have not been reported. We evaluated feasibility, acceptability, and costs after 1.5 years of IGRA use in San Francisco under routine program conditions. Patients seen at six community clinics serving homeless, immigrant, or injection-drug user (IDU) populations were routinely offered IGRA (Quantiferon-TB). Per guidelines, we excluded patients who were 17 years old, HIV-infected, immunocompromised, or pregnant. We reviewed medical records for IGRA results and completion of medical evaluation for TB, and at two clinics reviewed TB screening logs for instances of IGRA refusal or phlebotomy failure. Between November 1, 2003 and February 28, 2005, 4143 persons were evaluated by IGRA. 225(5%) specimens were not tested, and 89 (2%) were IGRA-indeterminate. Positive or negative IGRA results were available for 3829 (92%). Of 819 patients with positive IGRA results, 524 (64%) completed diagnostic evaluation within 30 days of their IGRA test date. Among 503 patients eligible for IGRA testing at two clinics, phlebotomy was refused by 33 (7%) and failed in 40 (8%). Including phlebotomy, laboratory, and personnel costs, IGRA use cost $33.67 per patient tested. IGRA implementation in a routine TB control program setting was feasible and acceptable among homeless, IDU, and immigrant patients in San Francisco, with results more frequently available than the historically described performance of TST. Laboratory-based diagnosis and surveillance for M. tuberculosis infection is now possible.

Table 3.2: An example of query-focused multi-document summarization, showing the question, the bottom-line summary and two of the source abstracts.

4 Preprocessing

Biomedical domain peculiarities Biomedical texts exhibit certain unique attributes that must be taken into account in the development of a summarization system. First, medical information arises in a wide range of document types [1]: electronic health records, scientific articles, semi-structured databases, X-ray images and even videos. Each document type presents very distinct characteristics that should be considered in the summarization process. We focus on scientific articles, which are mainly composed of text. Having knowledge about the article layout can be exploited to improve the summaries that are generated automatically [47]. Second, the specific nature of biomedical terminology makes it difficult to automatically process biomedical information [42]. Some of the discussed issues are as follows:

- Synonyms: The use of different terms to designate the same concept.
- Homonyms: The use of words/phrases with multiple meanings. For instance, the syntagms *coronary failure* and *heart attack* stand for the same concept, while the term *anaesthesia* may refer to either the *loss of sensation* or the procedure for *pain relief*.

- Neologisms: Newly coined words that are not likely to be found in a dictionary (e.g. the term *coumadinise* for the administration of coumadin).
- Elisions: The omission of words or sounds in a word or phrase. For example, *white count* which is understood by physicians as the *count of white blood cells*.
- Abbreviation: A shortened form of a word or phrase. For example, the use of *OCP* to refer to *oral contraceptive pills*.

Preprocessing steps In this study, if the abstract includes abbreviations, the abbreviations and their expansions are extracted. This information is then used to replace these shortened forms in the abstract body. For example, if the abbreviation defines *Autologous Bone Marrow Transplantation* as the expansion of *ABMT* for a particular abstract, this abbreviation would be replaced by *Autologous Bone Marrow Transplantation* anywhere else in the document body. If the abstract contains abbreviations and acronyms, but without any definition, the software[1] for abbreviation definition recognition presented in [21] is used. This software allows for the identification of abbreviations and their expansions in biomedical texts with an average precision of 95%. Abbreviations are then replaced by their expansions in the document body. Furthermore, we have used the stopword list included in nltk[2] extended with the PubMed stopwords[3] to remove the generic terms (e.g. prepositions and pronouns), which are not useful in our summarization process. We have also employed OpenNLP[4] to detect and split the sentences, and Stanford POS tagger [62] for tokening and part of speech tagging of each sentence.

5 Proposed Approach

5.1 Measuring Semantic Similarity using WordNet and UMLS

Problem Statement Many existing approaches to automatic summarization rely on comparing the similarity of two sentences in some ways. Most existing relatedness measures are based on knowledge sources such as concept hierarchies or ontologies. For general English text, research on measuring relatedness has relied on WordNet, a freely available database that can also be viewed as a semantic network. For clinical and biomedical vocabularies, they are compiled into UMLS, a large lexical and semantic database of medical terms maintained by the U.S. National Library of Medicine. Quantifying semantic relationships between linguistic items (terms) (e.g. synonymy, hypernymy, homonymy and co-occurrence relations) lies at the core of many NLP applications [46]. However, hard matching between words has long been an obstacle in identifying the relatedness of two sentences [66, 55]. The following examples illustrate such problems in general ([45]) and biomedical domains ([47]), respectively:

- General Domain:

  a1. Officers fired.

  a2. Several policemen terminated in corruption probe.

[1] Available at http://biotext.berkeley.edu/software.html
[2] http://nltk.org/
[3] http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/
[4] http://opennlp.sourceforge.net/

b1. Officers fired.

b2. Many injured during the police shooting incident.

Surface-based approaches that are merely based on string similarity cannot capture the similarity between any of the above pairs of sentences. In addition, a surface-based semantic similarity approach considers both *a1* and *b1* as being identical sentences, whereas different meanings of the verb fire are triggered in the two contexts [46]. To tackle this issue in computing WordNet-based sentence-to-query and sentence-to-sentence semantic similarities and any semantic ambiguity therein, we have adjusted and employed a unified approach proposed by [45]. To this end, WordNet 3.0 [15] repository has been used as our sense inventory.

- Biomedical Domain:

  1. Cerebrovascular diseases during pregnancy result from any of three major mechanisms: arterial infarction, haemorrhage or venous thrombosis.

  2. Brain vascular disorders during gestation result from any of three major mechanisms: arterial infarction, haemorrhage or venous thrombosis.

  Because the two sentences present different terms, surface-based approaches are unable to make use of the fact that they have exactly the same meaning. We have solved this problem by leveraging a UMLS-based approach dealing with concepts instead of terms, and with semantic relations instead of lexical or syntactical ones. This approach has previously been employed for query expansion [35].

In our work, the main requirement for computing semantic similarities on WordNet and UMLS is Semantic Signature. Pilehvar et al. [46] introduced semantic signature as a multinomial distribution generated from repeated random walks on WordNet. We utilize this concept to capture the semantic similarities on both WordNet and UMLS. Next, we briefly explain our journey to capture semantic similarities between sentences (note that in our work, a query is treated as a long single sentence).

Semantic Signature on WordNet To construct each semantic signature on WordNet, an iterative method for calculating Personalized PageRank has been used. The key assumption is that repeated random walks beginning at a sense (node) or a set of senses (seed nodes) in WordNet network can provide a frequency or multinomial distribution over all the senses in WordNet. A higher probability will then be assigned to senses that are frequently visited from the seeds. Consider an adjacency matrix *M* for the WordNet network, where edges connect senses according to the relations defined in WordNet (e.g. hypernymy and meronymy). A sense is further connected to all the other senses that appear in its disambiguated gloss.

The probability distribution for the starting location of the random walker in the network is denoted by $\vec{w}^{(0)}$. Given the set of senses *S* in a sentence, the probability mass of $\vec{w}^{(0)}$ is uniformly distributed across the senses $s_i \in S$, with the mass for all $s_i \notin S$ set to zero. The PageRank vector is then computed using Equation 5.1.

$$\vec{w}^{(t)} = (1-\alpha)M\vec{w}^{(t-1)} + \alpha\vec{w}^{(0)} \tag{5.1}$$

where at each iteration, the random walker may jump to any node $s_i \in S$ with probability $\alpha/|S|$. Following the standard convention, the value of α is set to 0.15. The number of iterations is also set to 30, which is sufficient for the distribution to converge. The resulting probability vector $\vec{w}^{(t)}$ is the semantic signature of the sentence, as it has aggregated its senses similarities over the entire graph. The UKB[1] implementation of Personalized PageRank has been used in this step.

WordNet-based Semantic Disambiguation In order to use a deeper modeling of linguistic items at the sense level, each word in a text has first to be analyzed and disambiguated into its intended sense. However, due to the inherent information shortage of sentences, traditional forms of word sense disambiguation (WSD) are hard to use. Therefore, we use an alignment-based sense disambiguation algorithm that has been presented in [46]. This algorithm leverages the content of the paired sentence in order to disambiguate each element.

Given two sentences, the semantic alignment procedure has been performed as follows: for each word type t_i in sentence S_1, assigns t_i to the sense that has the maximal similarity score to any sense of the word types in the compared sentence S_2. Let us consider the General Domain example:

P_{a1}. *officer*^3_n, $fire^4_v$
P_{a2}. $policeman^1_n, terminate^4_v, corruption^6_n, probe^1_n$

where P_i denotes the corresponding set of senses of sentence *i*. t_i denotes the i-th sense of a term t in WordNet with part of speech p. After alignment, among all possible pairings of all the senses of $fire_v$ to all the senses of all words in *a2*, the sense $fire^4_v$ (the employment termination sense) obtains the value $(Sim(fire^4_v, terminate^4_v) = 1)$, which is the maximal similarity value.

Semantic Signature on UMLS To construct each semantic signature on UMLS, we employ a graph-based algorithm to perform iterative random walks over the graph representation of the UMLS Metathesaurus. This algorithm has previously been utilized for query expansion [35]. The UMLS Metathesaurus contains a wide range of information about the relations between terms in the form of database tables. The MRREL table lists relations between concepts (i.e. *parent*, *can be qualified by*, and *related and possibly synonymous*) among others. Concepts in UMLS are considered as nodes (seeds), and the relations listed in the MRREL table as directed edges. No weights are used for the relations that are extracted from the MRREL table.

We have used the MetaMap program to map each sentence to concepts from the UMLS Metathesaurus and semantic types from the UMLS Semantic Network. A broad range of concepts from very generic UMLS semantic types are discarded in this step for two reasons: first, these generic concepts have already been considered in capturing WordNet-based semantic similarities; second, to reduce the size of UMLS graph, and consequently reduce the run time of iterative random walks. These semantic types are defined as *quantitative concept, qualitative concept, temporal concept, functional concept, idea or concept, intellectual product, mental process, spatial concept,* and *language* [47]. Thus, only concepts of the rest of semantic types are considered for constructing the semantic signature. Table 5.1 provides an example of mapping a sentence by MetaMap.

[1] http://ixa2.si.ehu.es/ukb/

Score	Concept	Semantic Type	Considered
862	No evidence of	Qualitative Concept	✘
593	Increase	Functional Concept	✘
593	Risk	Idea or Concept	✘
578	Major	Qualitative Concept	✘
744	Hemorrhage	Pathologic Function	✔
578	Result	Functional Concept	✘
578	Accidental Falls	Injury or Poisoning	✔
1000	Hospitalized Patients	Patient or Disabled Group	✔
966	Take	Health Care Activity	✔
1000	Warfarin	Pharmacologic Substance	✔

Table 5.1: MetaMap mapping for the sentence "*There is no evidence of increased risk for major bleeding as a result of falls in hospitalized patients taking warfarin.*"

Same as WordNet-based semantic signature, the UKB implementation of Personalized PageRank is utilized, but on UMLS. Consider an adjacency matrix *N* with all relations in MRREL, for the UMLS graph. The random walker starts in any of the concepts included in the sentence, and follows at random one of the relations to another concept. With certain probability, the random walker would restart in any of the concepts, and continue its walk. Finally, the number of visits to each concept in the graph would give an indication of how related that concept is to the sentence terms. The result is a probability distribution over UMLS concepts. The higher the probability for a concept, the more related it is to the given sentence.

The probability distribution for the starting location of the random walker in the network is denoted by $\vec{u}^{(0)}$. Released the set of MetaMap concepts *C* in a sentence, the probability mass of $\vec{u}^{(0)}$ is uniformly distributed across the concepts $c_i \in C$, with the mass for all $c_i \notin C$ set to zero. The PageRank vector is then computed using Equation 5.2.

$$\vec{u}^{(t)} = (1-\beta)N\vec{u}^{(t-1)} + \beta\vec{u}^{(0)} \tag{5.2}$$

where at each iteration, the random walker may jump to any node $c_i \in C$ with probability $\beta/|C|$. Following the standard convention, the value of β is set to 0.15. The number of iterations is also set to 30, which is sufficient for the distribution to converge. The resulting probability vector $\vec{u}^{(t)}$ is the semantic signature of the sentence on UMLS, as it has aggregated its concepts similarities over the entire graph.

UMLS-based semantic Disambiguation Using the built-in WSD module, MetaMap allows to disambiguate terms and return directly the relevant concept. For more clarity, we run MetaMap to find the UMLS concepts for the term *cold* (Figure 5.1). Normally, four concepts in MetaMap are assigned to this term. When WSD module is turned on, only one concept will be returned by considering the terms included in the given sentence. Then, a uniform probability distribution is assigned to the concepts found in each sentence. The rest of the nodes are initialized to zero.

Semantic Similarities at the Sentence Level For comparing pairs of semantic signatures at the sentence level, we have used Weighted Overlap (WO) approach proposed by [46]. WO first sorts the two signatures according to their values and then harmonically weights the overlaps between them. The weighting process is such that differences in the highest ranks are penalized more than differences in lower ranks (note that the first-ranked element has the highest rank). Using the knowledge source N (i.e. WordNet

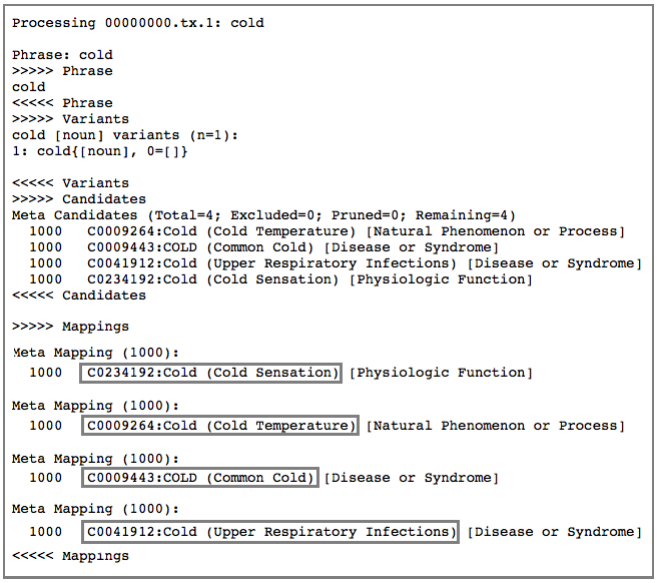


Figure 5.1: Screenshot of Mapping the term *cold* using MetaMap

or UMLS), WO calculates the semantic similarity (Sim_N) of two sentence signatures S_{N1} and S_{N2} as:

$$Sim_N(S_{N1}, S_{N2}) = \frac{\sum_{h \in H} (r_h(S_{N1}) + r_h(S_{N2}))^{-1}}{\sum_{i=1}^{|H|} (2i)^{-1}} \tag{5.3}$$

where *H* denotes the intersection of all senses/concepts with non-zero probability (dimension) in both signatures, and $r_h(S_{Nj})$ denotes the rank of the dimension *h* in the sorted signature S_{Nj}, where rank 1 denotes the highest rank. The denominator is also used as a normalization factor that guarantees a maximum value of one. The minimum value is zero and occurs when there is no overlap between the two signatures, i.e. $|H| = 0$.

To estimate the final semantic similarity score between two sentences, we have conducted a set of experiments using the WordNet-based semantic similarities (Sim_W), and/or UMLS-based semantic similarities (Sim_U), and obtained the best result while using both scores with different weights according to Equation 5.4.

$$Sim_{final}(S_1, S_2) = \mu \times Sim_U(S_{U1}, S_{U2}) + (1 - \mu) \times Sim_W(S_{W1}, S_{W2}) \tag{5.4}$$

where $Sim_U(S_{U1}, S_{U2})$ denotes the semantic similarity score between two sentence signatures on UMLS. The semantic similarity score between two sentence signatures on WordNet is also shown by $Sim_W(S_{W1}, S_{W2})$. Finally, the scaling factor μ was optimized on development data in our experiments and set to 0.6 to reach the best result (Section 6.2).

Next, using the achieved semantic similarity score for each pair of sentences, sentences which are less or not relevant to the clinical query will be pruned, and remained sentences will be clustered according to their relevance to each other.

5.2 Constructing Similarity Graph

In the field of generic query-sensitive summarization, qualified summary sentences should mainly meet the following typical demands: query-biased relevance [56], and biased information novelty and richness [64]. In this paper, we consider these criteria to develop an efficient framework for Query-focused EBM summarization. For being Query-biased relevant, summary sentences must overlap with the query in terms of topical content. Query-biased information novelty denotes that summary sentences need to be unique, as well as responding to the demands of the query. Finally, to acquire query-biased information richness, summary sentences should include as much important information as possible with respect to both the set of sentences and the query.

Query-biased Relevance To satisfy the query-biased relevance criterion, sentences are modeled as Similarity Graph - a weighted undirected graph on which each node represents a sentence and the edge weight carries the similarity of two sentences [55]. For more clarity, let $S = \{s_1, s_2, ..., s_n\}$, be a set of sentences, and $(S_{ij})_{i,j=1,...,N}$ be the similarity matrix in which each element indicates the similarity $S_{ij} \geq 0$ between two sentences S_i and S_j (pairwise similarity scores are already achieved in Section 5.1). Hence, the input query and the abstract sentences are considered as nodes on the graph, where we consider two kinds of edge for each node: (1) sentence-to-query similarity edge; (2) sentence-to-sentence similarity edge. The achieved similarity weight for each sentence-to-query and sentence-to-sentence relation is assigned to its corresponding edge in our similarity graph. This graph is partially depicted in Figure 5.2.

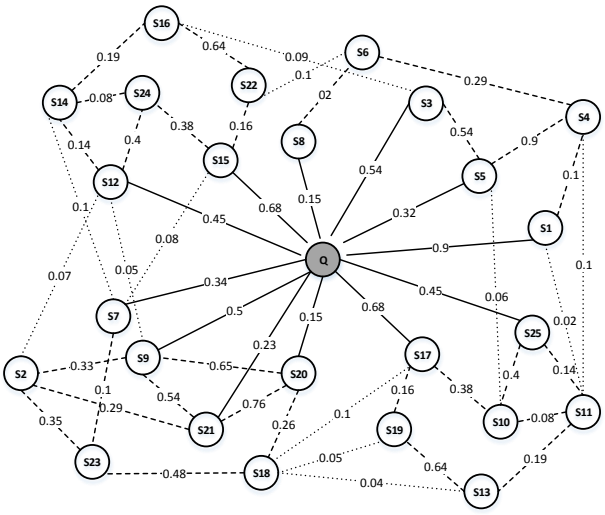


Figure 5.2: A partial view of the Similarity Graph

A sentence with a high sentence-to-query similarity score (direct query-biased sentence) is likely to include an answer to the query. Moreover, a sentence which may not be similar to the query, but still has a tight relation to a direct query-biased sentence, is also likely to include an answer. So, considering the combination of sentence-to-query

and sentence-to-sentence similarities, our model decides which sentences are relevant to the query, and should be kept for the further clustering step. To this end, we have employed a combination model [10]:

$$C(S_i|Q) = \gamma \times \frac{Sim_{final}(S_i, Q)}{\sum_{S_j \in A} Sim_{final}(S_j, Q)} + (1-\gamma) \times \sum_{S_k \in A} \frac{Sim_{final}(S_i, S_k)}{\sum_{S_j \in A} Sim_{final}(S_j, S_k)} \times C(S_k|Q) \quad (5.5)$$

where $C(S_i|Q)$ denotes the score of a sentence S_i given a query Q. A contains all sentences in the abstract set. The weighting parameter $0 \leq \gamma \leq 1$ is used to specify the relative contribution of two similarities: the similarity of a sentence to the query and similarity to the other sentences in the abstract set. Previous experiments [10] lead us to choose 0.4 as the best value of γ. The denominators in both terms are for normalization. $Sim_{final}(S_i, S_k)$ is the weight of the edge between two sentence nodes S_i, S_k. Likewise, $Sim_{final}(S_i, Q)$ is the weight of the edge connecting the sentence node S_i to the query node Q. Finally, sentences with $C \geq \delta$ (with the best empirical value of 0.5 for δ) are picked among the set of sentences. This step resulted in a subgraph comprising a set of the most query-relevant sentences to be clustered in the next step.

5.3 Clustering Relevant Sentences

The clustering problem from a graph perspective, is formulated as partitioning the graph into clusters such that the edges in the same cluster have high weights and the edges between different clusters have low weights. In this paper, we target hard clustering, where we partition nodes of the graph into non-overlapping clusters, i.e. let us partition S to a set of clusters $C = \{c_1, c_2, ..., c_n\}$ such that:

(1) $c_i \neq \phi \;\; for \;\; i \in \{1, ..., n\}$
(2) $c_i \cap c_j = \phi \;\; for \;\; i, j \in \{1, ..., n\} \;\; and \;\; i \neq j$
(3) $c_1 \cup ... \cup c_n = S$

The graph-based clustering algorithm we have used in this step, is the Chinese Whispers (CW) algorithm proposed by [4]. CW is a very basic - yet effective - parameter-free algorithm to partition the nodes of graphs in a bottom-up fashion. This algorithm is also a special case of Markov-Chain-Clustering [63], but time-linear in the number of edges. So, the power of CW lies in its capability of handling very large graphs in reasonable time. Algorithm 1 shows the adopted CW used in our work:

First, a distinct class is assigned to each node , and a clustering C containing the singleton clusters c_i is created (lines 1-4 of the algorithm). Then, a series of iterations is performed to merge the clusters (lines 5-11). Specifically, at each iteration the algorithm analyses each node s in random order and assigns it to the majority class among those associated with its neighbors. In other words, it assigns each node s to the class c that maximizes the sum of the weights of the edges s_i, s_j incident on s_j such that c is the class of s_i, according to Equation 5.6.

$$class(s_j) = \underset{c}{argmax} \sum_{\substack{\{s_i, s_j\} \in E(G) \\ s.t. class(s_i) = c}} Sim(s_i, s_j) \quad (5.6)$$

Algorithm 1 The Chinese Whispers (CW) Algorithm

Input: a graph $G = (S, E)$ to be clustered
Output: a clustering C of nodes in S

1: **For each** $s_i \in S$
2: $class(s_i) = i$
3: $C_i = \{s_i\}$
4: $C = \{c_i : i = 1, ..., |S|\}$
5: **repeat**
6: $C' = C$
7: **For each** $s_i \in S$, randomized order
8: $class(s_j) = \underset{c}{argmax} \sum_{\substack{\{s_i,s_j\}\in E(G) \\ s.t. class(s_i)=c}} Sim(s_i, s_j)$
9: **For each** i do $C_i = \{s_i \in S : class(s_i) = i\}$
10: $C = \{C_i : C_i \neq \phi\}$
11: **until** $C \neq C'$
12: **return** C

As soon as an iteration produces no change in the clustering (line 11), the algorithm stops and outputs the final clustering (line 12). The result of CW is a hard partitioning of the given graph into a number of clusters. Although it is possible to obtain a soft partitioning in CW, we prefer hard partitioning to keep the redundancy low.

Clustering Potential of the EBM Corpus As mentioned in [39], each query in the corpus is accompanied with multiple candidate replies. Since each candidate reply is referred to a set of abstracts, their released corpus could be utilized for the task of clustering. This ability is appreciated by an example shown in Table 5.2. However, we desire to consider a set of abstracts as a bag of sentences, pick the query-related sentences, and finally collect the relevant ones into a set of clusters. So, each cluster in our work is likely to include a set of sentences from different clusters defined in the corpus. Hence, we haven't used the clustering potential of corpus in this paper. Next, we build a word graph for each cluster, and generate one sentence as an abstractive summary of each cluster.

Question: What is the evaluation and treatment strategy for Raynaud's phenomenon?
Abstract IDs: 12814733, 12814733, 12324557, 11392916, 15865744, 10796398, 11508437
Resulting Clusters: Cluster1 → 12814733, 12814733, 12324557 Cluster2 → 11392916 Cluster3 → 15865744, 10796398, 11508437

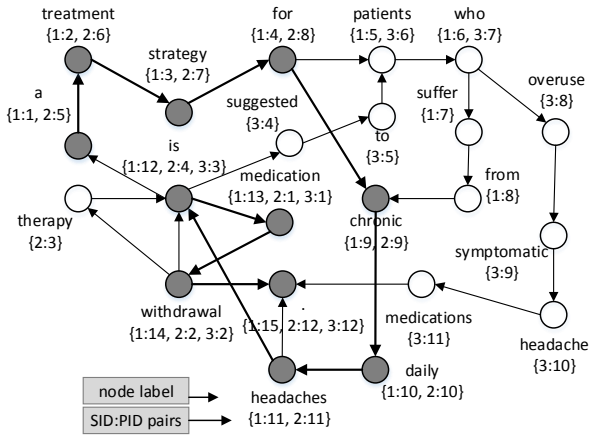
Table 5.2: An example of Clustering Potential of the Utilized Corpus

5.4 Abstractive Summarization of EBM

Biased Information Novelty In this section, we have built a word graph by iteratively adding sentences to it for each obtained cluster. This graph is an ordered pair $G = (V, E)$ comprising of a set of vervices or words, together with a set of directed edges, which shows the adjacency between corresponding nodes. The graph is first constructed by the first sentence and displays words in a sentence as a sequence of connected nodes. The first node is the start node and the last one is the end node. Words are added to the graph in three steps of the following order: (1) non-stopwords

for which no candidate exists in the graph; or for which an unambiguous mapping is possible; (2) non-stopwords for which there are either several possible candidates in the graph; or for which they occur more than once in the sentence; (3) stopwords. As mentioned in Section 4, for the last group, we used the stopword list included in nltk extended with the PubMed stopwords.

Where mapping in the graph is ambiguous (i.e. there are two or more nodes in the graph that refer to the same word/POS pair), we follow the instructions stated by [16]: the immediate context (the preceding and following words in the sentence, and the neighbouring nodes in the graph) or the frequency (i.e. the node which has words mapped to it) is used to select the candidate node. A new node is created only if there are no suitable candidates to be mapped to, in the graph. Conducting this step not only removes the redundancy, but also makes use of redundant parts to indicate the salient path (Figure 5.3). Edge weights are calculated using the weighting function defined in [16] (Equation 5.7).



Input:
SID:1. A treatment strategy for patients who suffer from chronic daily headaches is medication withdrawal.
SID:2. Medication withdrawal therapy is a treatment strategy for chronic daily headaches.
SID:3. Medication withdrawal is suggested to patients who overuse symptomatic headache medications.

Figure 5.3: An example of the Constructed Word Graph. Thick edges indicate salient paths.

$$W(e_{i,j}) = \frac{(freq(i) + freq(j))/\sum_{s \in S} diff(s,i,j)^{-1}}{freq(i) \times freq(j)} \tag{5.7}$$

where $freq(i)$ is the number of words mapped to the node i. The function $diff(s,i,j)$ refers to the distance between the offset positions of words i and j in sentence s.

Utilizing Synonymy To reduce the redundancy caused by existing synonym words in the sentences, we use the synsets in WordNet to identify synonym representative

candidate. For example, consider n different sentences containing words *biliary*, *bilious*, *tumor*, *tumour*, and *neoplasm*. The first two words, and the latter three ones are synonyms of each other. Assume each sentence contains one of these possible combinations (i.e. biliary tumor, biliary neoplasm, biliary tumour, bilious tumor, bilious neoplasm, bilious tumour). Without an appropriate synonym mapping based on a notion of synonymy, these several synonym nodes will be added to the word graph as separate nodes. We consider their frequency to pick one of them as the representative of its synonyms from the other sentences. The weight of the obtained node is computed by summing the frequency scores from the other nodes as shown in Figure 5.4. The main purpose of this modification is three fold: (i) the ambiguity of mapping nodes is reduced; (ii) the number of total possible paths (compression candidates) is decreased; and (iii) the weight of frequent similar words with different appearances in the content is better reflected by the notion of synonymy.

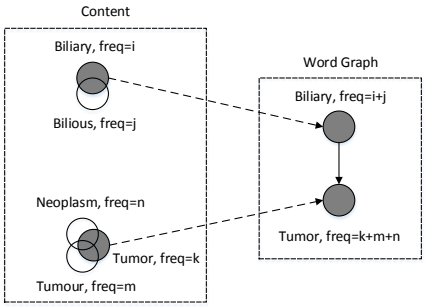


Figure 5.4: An example of Biomedical Synonym Mapping

The heuristic algorithm discussed in [6] is then used to find the k-shortest paths ($k = 50$ throughout our experiments) from start to end node in the graph. So, most of the potentially good candidates are kept and a decline in performance is prevented. Paths shorter than eight words or do not contain a verb are filtered before re-ranking. The remaining paths are re-ranked and the path that has the lightest average edge weight is eventually considered as the best compression.

Biased Information Richness To re-rank the compression candidates based on the information richness, important key-phrases have been exploited using the TextRank algorithm [36]. Hence, a word recommends other co-occurring words, and the strength of the recommendation is recursively computed based on the importance of the words making the recommendation. The score of a key-phrase *k* is computed by summing the salience of the words it contains, normalized with its length ($+1$) to favor longer n-grams. The paths are then re-ranked based on their key-phrases and the score of a compression candidate c is given by Equation 5.8.

$$Score_{Key}(c) = \frac{\sum_{i,j \in path(c)} W(e_{i,j})}{length(c) \times \sum_{k \in c} (\frac{\sum_{w \in k} TextRank(W)}{length(k)+1})} \tag{5.8}$$

Ensuring the Syntactic Structure Since our abstractive word graph generates new summary sentences, we need to ensure the grammatical structure of these newly constructed sentences. So, we build a part-of-speech based language model (POS-LM) to

re-rank the paths in our word graph [54]. The POS-LM assigns a score to each generated summary in terms of grammatical structure, and helps in identifying the most grammatical sentence among the k-best sentences. It estimates the probability of string of m POS tags by Equation 5.9 [40].

$$p(t_1^m) \propto \prod_{i-1}^{m} p(t_i|t_{i-n+1}^{i-1}) \tag{5.9}$$

where, *n* is the order of the language model, and t_i^j refers to the sub-sequence of tags from position *i* to *j*.

To build a POS-LM, we have employed the SRILM toolkit [59], which collects n-gram statistics from all n-grams occurring in the corpus, to build a single global language model. To train the POS-LM, we use Stanford POS tagger to annotate a large part ($\sim$100 M-words) of the BioMed Central full-text corpus for text mining research[2] that contains a large number ($\sim$ 290914) of biomedical articles. Then, we remove all words from the pairs of words/POS in the POS annotated corpus. The candidate sentences also need to be annotated with POS tags. So, the score of each summary is estimated by the language model, based on its sequence of POS tags. Since factors like POS tags, are less sparse than surface forms, it is possible to create higher order language models for these factors. This may encourage more syntactically correct output [30]. Thus, for our framework, we use 7-gram language modeling based on part of speech tagging to re-rank the k-best sentences generated by the word graph. To re-rank the obtained paths, POS-LM gives the perplexity score ($Score_{LM}$), which is the geometric average of $1/probability$ of each sentence, normalized by the number of words. So, $Score_{LM}$ for each sequence of POS in the k-best sentences is computed by Equation 5.10.

$$Score_{LM}(c) = 10^{\frac{\log\, prob(c)}{|word|}} \tag{5.10}$$

where $prob(c)$ is the probability of summary C including $|word|$ number of words, computed by the 7-gram POS-LM.

A unity-based normalization is then used to bring the values of $Score_{Key}(c)$ in Equation 5.8, and the score of POS-LM into the range [0, 1]. The score of each summary is finally given by Equation 5.11.

$$Score_{final}(c) = \eta \times Score_{Key}(c) + (1 - \eta) \times Score_{LM}(c) \tag{5.11}$$

The scaling factor η was optimized on development data in our experiments and set to 0.4 (Section 6.2). Syntactic analysis of the generated sentences is also explored in Section 6.2. Hence, the most grammatical candidate among the candidates contain the most important phrases, has been selected as the summary for each cluster.

All automatic summaries were generated by selecting sentences until the summary is 30% of the original document size [47]. This choice of the summary size is based on the well-accepted heuristic that a summary should be between 15% and 35% of the size of the source text. Considering this convention, we pick a number of three summary sentences (based on their sentence-to-query similarity scores) to answer the corresponding clinical query.

[2]http://old.biomedcentral.com/about/datamining

6 Evaluation

6.1 Evaluation Metrics

The evaluation of automatically generated summaries is a critical issue due to the subjectivity in deciding of what the evaluation criteria should be [48]. The evaluation process may be performed manually, and require human judges to decide whether or not a summary is of good quality. So, manual evaluation is very costly and time-consuming. Besides, to objectively judge a summary has been proven difficult, as humans often disagree on what exactly makes a summary of good quality [29]. Given that, for our summarization framework, the generated summaries are assessed automatically through version 2.0 [1] of ROUGE [31] over the released EBM corpus by [39].

ROUGE is a commonly used evaluation method to measure the summary quality by counting the overlapping units between system-generated summaries and human-written reference/gold summaries. ROUGE measures the concordance of candidate and reference summaries by determining *n*-gram, word sequence, and word pair matches. The ROUGE metrics produce a value in [0,1], where higher values are preferred, as they indicate a greater content overlap between the generated summary and reference summaries. We have used ROUGE F-measure for unigram, bigrams, and SU4 (skip-bigram with maximum gap length 4) to evaluate the generated summaries. The bottom-line answers in the EBM corpus have also been used as the reference summaries.

An important drawback of ROUGE metrics is that they use lexical matching instead of semantic matching. Therefore, generated summaries that are worded differently but carry the same semantic information may be assigned different ROUGE scores [47]. In contrast, the main advantages of ROUGE are its simplicity and its high correlation with the human judges, based on the results reported in the previous DUC conferences [31].

6.2 Experiments

To investigate the effectiveness of our abstractive summarization framework for EBM, we compare our framework with *FastSum* [53], and a research prototype *LexRank* [14]. FastSum is a fast query-focused multi-document summarization system based only on word frequency features of topics, documents, and clusters. Each sentence is ranked based on a linear function of scores using a variety of frequency measures. A regression SVM is also used to learn weights of the features. LexRank is a topic-oriented generic summarizer that focuses on multi-document extractive text summarization, and extracts the information in the text that is related to the user specified topic. This prototype has outperformed both centroid-based methods and other systems participating in DUC in most of the cases [14]. Comparison with LexRank will allow us to evaluate whether semantic information provides benefits over merely lexical information in graph-based summarization approaches.

In addition, we pick the first and last third sentences of each set of abstracts related to a clinical query, so called (*first part*, and *last part*). We also consider all sentences included in the abstracts related to a clinical query as *whole part*. Afterwards, included sentences in each of these three parts are considered as the input bag of sentences for the following baselines:

- Head Baseline: This baseline is used in a variety of summarization applications, specifically in the news summarization area. In our work, this baseline generates

[1] http://kavita-ganesan.com/content/rouge-2.0

summaries by unintentionally selecting three sentences from the *first part*.

- Random Baseline: Randomly selects three sentences from the *whole part*.
- Tail Baseline: The last sentences in the medical abstracts usually provide conclusions. Hence, this has been used as a baseline for summarization of biomedical texts [12]. In our work, this baseline generates summaries by selecting three sentences at random from the *last part*.

Furthermore, the effectiveness of the abstractive summarizer of our framework, along with the re-ranking algorithms, is also studied using the following experiments. We keep consistency for our framework algorithm except to omit the graph-based clustering and the word graph, and converting our abstractive framework to the ranking-based extractive approach (Proposed-Ext). For more clarity, we have conducted only two first components of our framework, which are capturing semantic WordNet and UMLS-based sentence-to-query and sentence-to-sentence similarities, and also sentence filtering step to achieve the most query-relevant sentences. The average performance of the baseline systems and the proposed framework in terms of ROUGE scores are shown in Figure 6.1, and the data is provided in Table 6.1.

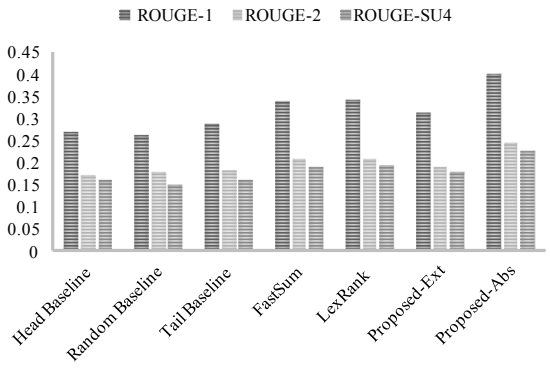


Figure 6.1: Average scores by ROUGE metrics over the EBM corpus

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Head Baseline	0.2710	0.1723	0.1593
Random Baseline	0.2623	0.1801	0.1509
Tail Baseline	0.2866	0.1834	0.1607
FastSum	0.3382	0.2081	0.188
LexRank	0.3407	0.2069	0.1938
Proposed-Ext	0.3142	0.1911	0.1806
Proposed-Abs	**0.3985**	**0.2450**	**0.2259**

Table 6.1: Average scores by ROUGE metrics over the EBM corpus

The statistics point out the effectiveness of our abstractive framework over the compared systems on all evaluation metrics. Hence, the overall results support our hypothesis that query-based abstractive summarization using the underlying textual semantic similarities based on both WordNet and UMLS knowledge sources results in significantly better performance. Besides, the results achieved by Proposed-Ext on the

EBM corpus still show some improvements over some of the baseline systems. The main reason may be capturing both WordNet and UMLS-based semantic similarities, which help to select the most query-relevant sentences among the set of biomedical abstracts. Finally, considering the results obtained by *Tail Baseline*, it has been realized that the last part of each abstract is more likely to be included in the summary. Table 6.2 shows an example of a summary generated by human (gold), our abstractive framework (Proposed-Abs), and the extractive LexRank.

Question: Are major bleeding events from falls more likely in patients on warfarin?
Gold Summary: There is no evidence of increased risk for major bleeding as a result of falls in hospitalized patients taking warfarin. [*PubMed IDs: 7668955, 15638939*]
Proposed-Abs Summary One study found no difference in major bleeding complications between patients taking anticoagulation therapy with not taking. Criteria for taking warfarin were not reported. Prescribing warfarin for patients judged less likely to fall.
LexRank Summary No major hemorrhagic complications were seen following 131 falls in the anticoagulation group (93 patients) and 269 falls in the group not on anticoagulation (175 patients). The study was limited because most falls were from a seated position or partially controlled by an attendant. Major hemorrhage was defined as bruising or cuts requiring immediate attention from a physician.

Table 6.2: An example of different summaries: Gold summary; Proposed-Abs summary; and LexRank summary.

Standard Deviation of ROUGE Scores Since Table 6.1 shows the average results, an important research question that immediately arises is how much the ROUGE scores differ across the abstracts. Hence, the standard deviation of different ROUGE scores for the summaries generated by Proposed-Abs are shown in Table 6.3.

Metric	ROUGE-1	ROUGE-2	ROUGE-SU4
Standard Deviation	0.02104	0.03250	0.03079

Table 6.3: Standard deviation of ROUGE scores for the summaries generated by Proposed-Abs

Exploring Scaling Factors In our work, two free parameters are defined: *Scaling Factor 1* (μ in Equation 5.4 - *measuring semantic similarities using WordNet and UMLS*), and *Scaling Factor 2* (η in Equation 5.11 - *The final re-ranking score of each generated summary sentence*). We randomly selected 30% of the EBM corpus as a development set to tune these parameters. Figure 6.2 shows the results obtained by ROUGE-1 F-Measure, using different values for μ and η. The best results are obtained using $\mu = 0.6$, and $\eta = 0.4$. Performance deteriorates when the UMLS portion in measuring semantic similarities is less or more than 0.6. On the other hand, when contribution of *TextRank* score for each generated summary sentence is whatever except 0.4, the performance gradually decreases. The lowest performance is obtained when *TextRank* score is ignored in re-ranking the generated summary sentences, and also when UMLS semantic signature occupies 0.9 of whole 1.0 value of final semantic similarity measure. This demonstrates the importance of using both WordNet and UMLS to capture the semantic similarities.

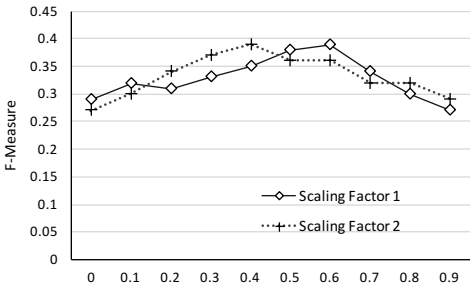


Figure 6.2: Accuracy results when exploring different values for *Scaling Factor 1* (μ in Equation 5.4), and *Scaling Factor 2* (η in Equation 5.11) on the development set.

Syntactic Analysis of the Generated Sentences Finally, we have analyzed a random selected part of the generated summaries in terms of syntactic structure, using version 5.3.7 of Link Grammar Parser[2] (Figure 6.3). This parser is a syntactic analyzer of English language developed at the Carnegie Mellon University [61, 58]. Having received a sentence, the system attributes it with a syntactic structure which consists of a set of marked links connecting the pairs of words. It includes approximately 60000 dictionary forms, and can skip a part of a sentence it cannot understand and define some structure for the rest of the sentence. It is capable of processing an unknown lexicon and doing reasonable assumptions about the syntactic category of unknown words based on the context and writing. The parser contains data about various names, numerical expressions, and punctuation marks.

We have performed a random selection of generated summary sentences among the set of high ranked ones by the POS-LM, to syntactically analyze them using the Link Grammar, and consequently show the effectiveness of our grammar-enhanced re-ranking step. The parser gives a constituent representation of the sentence, labeling noun phrases, verb phrases, clauses, etc.. The constituent representation is derived from the linkage. The parser does not consider a sentence to be "grammatical", just because it finds a valid linkage for that sentence. The linkage must satisfy a post-processing phase. The parser indicates the post-processing status with messages like "Found 2 linkages (1 with no P.P. violations)". If all of the linkages at one stage have post processing violations, the parser continues looking for a satisfactory linkage in the next phase.

If there is more than one satisfactory linkage, the parser orders them according to certain simple heuristics. The cost vector determines the ordering used. This vector has three components. The first component (most significant in the ordering) is the total cost of all the usages of words in the linkage. The dictionary assigns different costs to the usages of a word; while most usages have cost nothing, some have non-zero cost. The second component has to do with the relative size of components combined by conjunctions. The third component is the total length of all links in the linkage. Figure 6.3 demonstrates this process for a sample generated sentence *"A treatment strategy for chronic daily headaches is medication withdrawal."*. The Link Grammar finds two complete linkages with no p.p violations, which indicates that this newly generated summary sentence is grammatically correct. This analysis shows about 85% precision over the syntactic structure of summary sentences. In details, as shown in Figure 6.4, among 600 random-selected summary sentences, 512 sentences have been shown to

[2]http://www.abisource.com/projects/link-grammar/

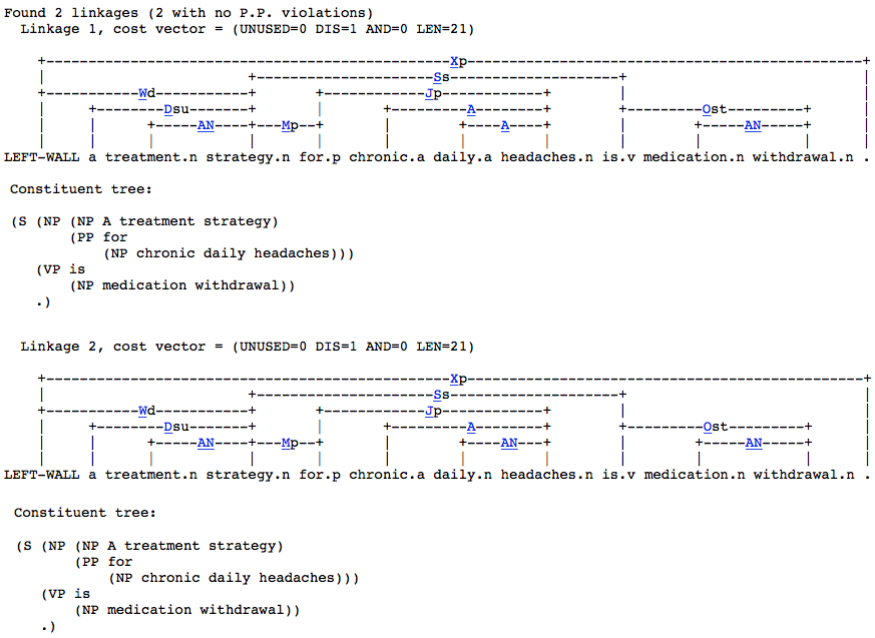


Figure 6.3: An example of using Link Grammar Parser for the syntactic analysis of a sample generated sentence *"A treatment strategy for chronic daily headaches is medication withdrawal."*

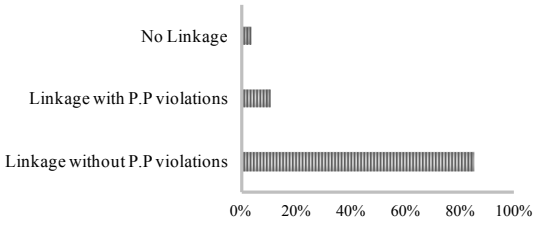


Figure 6.4: Syntactic analysis of a number of 600 generated summary sentences using Link Grammar Parser

have at least one complete linkage with no P.P violations, 64 sentences (11%) have a number of linkages but with some P.P violations, and finally, Link Grammar parser cannot find any linkage for 24 sentences (4%).

7 Conclusions

We have presented an effective approach for summarizing biomedical texts. Given a clinical query, our approach generates a well-organized, informative summary from a set of related biomedical abstracts through: (1) repetitive random walks on WordNet and UMLS to capture semantic similarities between sentences and the input query; (2) filtering out less query-relevant sentences; (3) clustering the remaining relevant

sentences using a graph-based clustering algorithm; (4) abstractive summarization of the clusters through a word graph-based approach, which considers the important key-phrases, along with the syntactic structure of the generated summaries. Based on an automatic evaluation (via ROUGE metrics) using an evidence-based medicine corpus, our framework outperforms the two competitive systems. Three different baselines for sentence selection have also been used, each aiming to construct a different type of summary according to the type of information in various parts of the source. It has been found that, the last part of each abstract is more likely to be included in the summary.

Our approach has significantly satisfied query-biased relevance, biased information novelty, and biased information richness. We have tackled the main issue faced by state-of-the-art biomedical summarizers (i.e. decline in summarization efficiency due to the poor UMLS coverage of general concepts in the documents to be summarized) [47]. This issue is addressed by using WordNet to represent the layman knowledge, and UMLS to represent the professional knowledge. We believe that this approach can bridge the knowledge and language gaps in biomedical summarizers.

A medium sized corpus (which is the only available corpus for evidence-based biomedical summarization) was used in our experiments. Hence, for some features, there was not enough data available for the generation of statistics. For example, the corpus only contains a few samples for some of the question types, e.g., History and Device. Having a larger corpus would make the statistics associated with sparse data more reliable. Therefore, our ongoing work includes constructing a larger corpus for evidence-based biomedical summarization.

Bibliography

[1] Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

[2] Alan R Aronson. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*, pages 1–26, 2006.

[3] Sofia J Athenikos and Hyoil Han. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24, 2010.

[4] Chris Biemann. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pages 73–80. Association for Computational Linguistics, 2006.

[5] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

[6] Florian Boudin and Emmanuel Morin. Keyphrase extraction for n-best reranking in multi-sentence compression. In *North American Chapter of the Association for Computational Linguistics*, 2013.

[7] Ari D Brooks and Isabel Sulimanoff. Evidence-based oncology project. *Surgical Oncology Clinics of North America*, 11(1):3–10, 2002.

[8] Anita Burgun and Olivier Bodenreider. Comparing terms, concepts and semantic classes in wordnet and the unified medical language system. In *Proceedings of*

the NAACL2001 Workshop,WordNet and Other Lexical Resources: Applications, Extensions and Customizations, pages 77–82, 2001.

[9] Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–288, 2011.

[10] Yllias Chali, Sadid A Hasan, and Shafiq R Joty. Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels. *Information Processing & Management*, 47(6):843–855, 2011.

[11] Herma CH Coumou and Frans J Meijman. How do primary care physicians seek answers to clinical questions? a literature review. *Journal of the Medical Library Association*, 94(1):55, 2006.

[12] Dina Demner-Fushman and Jimmy Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[13] John W Ely, Jerome A Osheroff, Mark H Ebell, M Lee Chambliss, Daniel C Vinson, James J Stevermer, and Eric A Pifer. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *British Medical Journal*, 324(7339):710, 2002.

[14] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479, 2004.

[15] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

[16] Katja Filippova. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 322–330. Association for Computational Linguistics, 2010.

[17] Marcelo Fiszman, Thomas C Rindflesch, and Halil Kilicoglu. Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 76–83. Association for Computational Linguistics, 2004.

[18] Clifford W Gay, Mehmet Kayaalp, and Alan R Aronson. Semi-automatic indexing of full text biomedical articles. In *American Medical Informatics Association*, 2005.

[19] Evidence-Based Medicine Working Group et al. Evidence-based medicine. a new approach to teaching the practice of medicine. *Journal of the American Medical Association*, 268(17):2420, 1992.

[20] Allan Hanbury. Medical information retrieval: an instance of domain-specific search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1191–1192. Association for Computing Machinery, 2012.

[21] MAAS Schwartz Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. 2003.

[22] William R Hersh, M Katherine Crabtree, David H Hickam, Lynetta Sacherek, Charles P Friedman, Patricia Tidmarsh, Craig Mosbaek, and Dale Kraemer. Factors associated with success in searching medline and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association*, 9(3):283–293, 2002.

[23] Deirdre Hogan. Empirical measurements of lexical similarity in noun phrase conjuncts. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 149–152. Association for Computational Linguistics, 2007.

[24] Dimitar Hristovski, Dejan Dinevski, Andrej Kastrin, and Thomas C Rindflesch. Biomedical question answering using semantic relations. *BMC bioinformatics*, 16(1):1, 2015.

[25] Kuo-Chuan Huang, James Geller, Michael Halper, Yehoshua Perl, and Junchuan Xu. Using wordnet synonym substitution to enhance umls source integration. *Artificial Intelligence in Medicine*, 46(2):97–109, 2009.

[26] Lawrence Hunter and K Bretonnel Cohen. Biomedical language processing: what's beyond pubmed? *Molecular cell*, 21(5):589–594, 2006.

[27] MTI Improving. Identification of important text in full text articles using summarization.

[28] K Sparck Jones et al. Automatic summarizing: factors and directions. *Advances in Automatic Text Summarization*, pages 1–12, 1999.

[29] Karen Sparck Jones and Julia R Galliers. *Evaluating natural language processing systems: An analysis and review*, volume 1083. Springer Science & Business Media, 1995.

[30] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[31] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, volume 8, 2004.

[32] Lauren A Maggio MS LIS, Ryan M Steinberg MSI, and Laura Moorhead. Access of primary and secondary literature by health personnel in an academic health center: implications for open access*. *Journal of the Medical Library Association*, 101(3):205, 2013.

[33] Beier Lu. *Health Query Expansion Using WordNet and UMLS*. PhD thesis, Applied Sciences:, 2015.

[34] Iain J Marshall, Joël Kuiper, and Byron C Wallace. Automating risk of bias assessment for clinical trials. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1406–1412, 2015.

[35] David Martinez, Arantxa Otegi, Aitor Soroa, and Eneko Agirre. Improving search over electronic health records using umls-based query expansion through random walks. *Journal of Biomedical Informatics*, 51:100–106, 2014.

[36] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004.

[37] Muhidin A Mohamed and Mourad Oussalah. Similarity-based query-focused multi-document summarization using crowdsourced and manually-built lexical-semantic resources. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, volume 2, pages 80–87. IEEE, 2015.

[38] Diego Mollá, Maria Elena Santiago-Martinez, et al. Development of a corpus for evidence based medicine summarisation. 2011.

[39] Diego Mollá, María Elena Santiago-Martínez, Abeed Sarker, and Cécile Paris. A corpus for research in text processing for evidence based medicine. *Language Resources and Evaluation*, pages 1–23, 2015.

[40] Christof Monz. Statistical machine translation with local language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 869–879. Association for Computational Linguistics, 2011.

[41] Fleur Mougin, Anita Burgun, and Olivier Bodenreider. Using wordnet to improve the mapping of data elements to umls for data sources integration. In *American Medical Informatics Association Annual Symposium Proceedings*, volume 2006, page 574. American Medical Informatics Association, 2006.

[42] PM Nadkarni. E-medicine-information retrieval in medicine: Overview and applications. 2000.

[43] Yun Niu, Xiaodan Zhu, and Graeme Hirst. Using outcome polarity in sentence extraction for medical question-answering. In *American Medical Informatics Association*, 2006.

[44] Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. Analysis of polarity information in medical text. In *American Medical Informatics Association*, 2005.

[45] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Association for Computational Linguistics*, pages 1341–1351, 2013.

[46] Mohammad Taher Pilehvar and Roberto Navigli. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128, 2015.

[47] Laura Plaza, Alberto Díaz, and Pablo Gervás. A semantic graph-based approach to biomedical summarisation. *Artificial Intelligence in Medicine*, 53(1):1–14, 2011.

[48] Dragomir R Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, and Elliott Drabek. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 375–382. Association for Computational Linguistics, 2003.

[49] Lawrence H Reeve, Hyoil Han, and Ari D Brooks. The use of domain-specific concepts in biomedical text summarization. *Information Processing & Management*, 43(6):1765–1776, 2007.

[50] David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72, 1996.

[51] Abeed Sarker, Diego Mollá, and Cécile Paris. Automatic evidence quality prediction to support evidence-based decision making. *Artificial Intelligence in Medicine*, 64(2):89–103, 2015.

[52] Abeed Sarker, Diego Mollá, and Cecile Paris. Query-oriented evidence extraction to support evidence-based medicine practice. *Journal of Biomedical Informatics*, 59:169–184, 2016.

[53] Frank Schilder and Ravikumar Kondadadi. Fastsum: fast and accurate query-based multi-document summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 205–208. Association for Computational Linguistics, 2008.

[54] Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. On improving informativity and grammaticality for multi-sentence compression. *arXiv preprint arXiv:1605.02150*, 2016.

[55] Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. A query-based summarization service from multiple news sources. In *Services Computing, 2016 IEEE International Conference*. IEEE, 2016.

[56] Chao Shen and Tao Li. Learning to rank for query-focused multi-document summarization. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 626–634. IEEE, 2011.

[57] 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 *Advances in Artificial Intelligence*, pages 284–295. Springer, 2007.

[58] Daniel DK Sleator and Davy Temperley. Parsing english with a link grammar. 1991.

[59] Andreas Stolcke et al. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*, volume 2002, page 2002, 2002.

[60] Rodney L Summerscales. *Automatic summarization of clinical abstracts for evidence-based medicine*. PhD thesis, Illinois Institute of Technology, 2013.

[61] Davy Temperley, John Lafferty, and Daniel Sleator. Link grammar parser. *: http://www.link.cs.cmu.edu/link*, 1995.

[62] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[63] Stijn Van Dongen. A cluster algorithm for graphs. *Report-Information systems*, (10):1–40, 2000.

[64] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Manifold-ranking based topic-focused multi-document summarization. In *International Joint Conference on Artificial Intelligence*, volume 7, pages 2903–2908, 2007.

[65] Wang Weiming, Dawei Hu, Min Feng, and Liu Wenyin. Automatic clinical question answering based on umls relations. In *Third International Conference on Semantics, Knowledge and Grid*, pages 495–498, 2007.

[66] Wenpeng Yin, Lifu Huang, Yulong Pei, and Lian'en Huang. Relationlistwise for query-focused multi-document summarization. In *International Conference on Computational Linguistics*, pages 2961–2976, 2012.

[67] Hong Yu and YongGang Cao. Automatically extracting information needs from ad hoc clinical questions. In *American Medical Informatics Association*, 2008.