Appraising UMLS Coverage for Summarizing Medical Evidence

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, Fang Chen
University of New South Wales, Sydney, Australia
Data61 CSIRO, Australia
{elahehs,mohammade,wong,fang}@cse.unsw.edu.au

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. This paper proposes an effective summarizer for medical evidence by utilizing both UMLS and WordNet. Given a clinical query and a set of relevant abstracts, we aim to generate a fluent, well-organized, and compact summary that answers the query. 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 approach is demonstrated by conducting a set of experiments over a specialized evidence-based medicine (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* (Sackett et al., 1996). Evidence-based medical practice heavily relies on research evidence, rather than intuition, unsystematic clinical experience, or pathologic rationale (Group and Others, 1992). However, searching through and evaluating primary medical literature is extremely time consuming (Sarker et al., 2015). Even targeted searches tend to return a large set of relevant documents, and not summaries or answers to the queries. 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 (Reeve et al., 2007).

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. They usually work with a simple representation of the summary comprising of information that can be directly extracted from the document, such as terms, phrases, or sentences (Mihalcea and Tarau, 2004). However, recent studies (e.g. (Fiszman et al., 2004)) 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, etc.) (Plaza et al., 2011).

While a query is asked in the field of biomedicine, one of the 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. 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, etc.) (Plaza et al., 2011). Therefore, particular characteristics and the type of

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 513–524, Osaka, Japan, December 11-17 2016.

biomedical documents are required to be exploited by the summarization systems. To this end, promising domain specific NLP techniques have been efficiently employed to release a repository of biomedical vocabularies named the Unified Medical Language System (UMLS[1]) (Bodenreider, 2004). 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[2] (Huang et al., 2009). Moreover, utilizing WordNet to complement the UMLS coverage is challenging due to their different structures, natures, terms, and sizes.

This challenge has motivated us to provide a deeper analysis of biomedical texts by keeping an eye on the biomedical peculiarities. 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. In details, our approach comprises different components: capturing underlying sentence-to-query and sentence-to-sentence semantic similarities using WordNet and UMLS; ranking and filtering sentences considering their similarity scores to the clinical query; clustering sentences by their relevance to each other; generating new summary sentences through a word graph representation by ensuring their importance and syntactic structure.

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

2 Background

Text summarization is the process of automatically creating a compressed version of a given text. A summary can either be query-focused (biased to a user query), or generic (conveying the document gist). In traditional query-focused summarization systems, lexical similarity measures are 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 question answering perspective (Yu and Cao, 2008), and some others have modeled summarization as a sentence classification problem (Cao et al., 2011). A machine learning classifier trained on a small dataset is employed in another study (Demner-Fushman and Lin, 2007) to select the summary sentences. Another summarization system (Cao et al., 2011) 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. More advanced summarization techniques such as LexRank (Erkan and Radev, 2004) incorporate graph-based methods. LexRank assumes a fully connected and undirected graph for the set of documents to be summarized.

Among the researches performed in the text summarization area, many studies (e.g. (Coumou and Meijman, 2006)) have also explored the obstacles associated with evidence-based medicine practice in the absence of pre-existing systematic reviews. When primary care physicians seek answers to clinical problems, the time required to search, evaluate, and synthesize evidence has been known as a critical factor (Sarker et al., 2016). 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 (Hersh et al., 2002)). Numerous IR approaches have already been proposed to address the search-related needs of practitioners (Hanbury, 2012). However, post-retrieval techniques (e.g. (Sarker et al., 2016)) to perform query-oriented summarization are still scarce. The complicated nature of biomedical texts and the limited amount of suitable annotated data for the task of summarization are the main reasons that raise various difficulties in progress (Athenikos and Han, 2010; Sarker et al., 2016).

To overcome the lack of incorporation of domain-specific information, UMLS came to play, and has proved to be a useful knowledge source for summarization in biomedical domain (Reeve et al., 2007).

[1]Developed by the U.S. National Library of Medicine (available at http://www.nlm.nih.gov/research/umls/)

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

However, a decline is found in the performance of the 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 (Plaza et al., 2011). To compensate this deficiency, a question-oriented extractive system for biomedical multi-document summarization (i.e. (Shi et al., 2007)), utilized WordNet as a general-purpose lexicon to capture the concepts not covered by UMLS. They constructed a graph containing ontological concepts (general ones from WordNet, and specific ones from UMLS), name entities, and noun phrases. Our work differs in intent, and explores the utility of graph representation of both domain-independent (WordNet) and domain-specific (UMLS) lexicons for incorporating underlying textual semantic similarities as the main basis of an efficient biomedical summarizer. Next, we discuss the utilized data and the preprocessing steps.

3 Data and Preprocessing

To develop, test, and evaluate our approach, we employed the evidence-based medicine (EBM) corpus[3] gathered and annotated by (Mollá et al., 2015), which is the only available corpus for the task of EBM text summarization. This corpus is sourced from the Clinical Inquiries section of the Journal of Family Practice[4], and consists of 456 clinical queries, with 1396 bottom-line multi-document summaries (i.e. evidence-based answers). The total number of associated single-document evidence-based summaries is 3036, which are generated from 2908 unique articles. Table 1 lists the properties of this corpus. 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.

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

Table 1: Information about the EBM Corpus

The specific nature of the biomedical terminology makes it difficult to automatically process biomedical information (Nadkarni, 2000). One of these difficulties is caused by abbreviations (e.g. the use of *OCP* instead of *oral contraceptive pills*). In our approach, if the abstract includes abbreviations, their expansions are used to replace these shortened forms in the abstract body. If the abstract contains abbreviations and acronyms, but without any definition, the software[5] for abbreviation recognition and definition presented in (Hearst, 2003) is used. To remove the stopwords, we utilized the stopword list included in nltk[6] extended with the PubMed stopwords[7]. We also employed OpenNLP[8] to detect and split the sentences, and Stanford POS tagger (Toutanova et al., 2003) for tokenizing and part of speech tagging of each sentence.

4 The Proposed Approach

4.1 Measuring Semantic Similarity using WordNet and UMLS

Automatic summarization approaches rely on similarity comparison of sentences. For general English text, research on measuring relatedness has relied on WordNet, and for clinical and biomedical vocabularies, they are compiled into UMLS. Quantifying semantic relationships between linguistic terms lies at the core of many NLP applications (Pilehvar and Navigli, 2015). However, hard matching between words has long been an obstacle in identifying the relatedness of two sentences (ShafieiBavani et al., 2016b). We tackle this issue by dealing with concepts instead of terms, and with semantic relations

[3] Available at: http://sourceforge.net/projects/ebmsumcorpus
[4] http://www.jfponline.com/articles/clinical-inquiries.html
[5] Available at http://biotext.berkeley.edu/software.html
[6] http://nltk.org/
[7] http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/
[8] http://opennlp.sourceforge.net/

instead of lexical or syntactical ones. In our approach, the main requirement for computing semantic similarities on WordNet and UMLS is Semantic Signature, which is firstly introduced as a multinomial distribution generated from repeated random walks on WordNet (Pilehvar and Navigli, 2015). We utilize this concept to capture the semantic similarities on both WordNet and UMLS. 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, we make use of WordNet 3.0 (Fellbaum, 1998) repository. We also employ an alignment-based sense disambiguation algorithm presented in (Pilehvar and Navigli, 2015) to disambiguate each word. This algorithm leverages the content of the paired sentence in order to disambiguate each element. Afterwards, 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 the 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). 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 1.

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

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[9] implementation of Personalized PageRank has been used in this step.

Semantic Signature on UMLS Each semantic signature on UMLS is constructed by performing iterative random walks over the graph representation of version 2015AB of the UMLS Metathesaurus. This algorithm has previously been utilized for query expansion (Martinez et al., 2014). Metathesaurus, Semantic Network and SPECIALIST Lexicon are three major components of UMLS. Our approach focuses on the UMLS Metathesaurus, which contains a wide range of information about the relations between terms in the form of database tables. Among them, MRREL table lists different relations between concepts (i.e. *parent*, *can be qualified by*, and *related and possibly synonymous*). We consider the UMLS concepts as nodes (seeds), and the relations listed in MRREL table as directed edges.

Besides, we employ version 2016 of the MetaMap[10] program to map each sentence to concepts from the UMLS Metathesaurus and semantic types from the UMLS Semantic Network. Using the built-in WSD module, MetaMap allows to disambiguate terms, and returns directly the relevant concept. A broad range of concepts from very generic UMLS semantic types, that have already been considered in capturing WordNet-based semantic similarities, are discarded in this step. 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* (Plaza et al., 2011). Thus, only concepts of the rest of semantic types are considered for constructing the semantic signature. Table 2 provides an example of mapping a sentence by MetaMap. Same as WordNet-based semantic signature, the UKB implementation of Personalized PageRank is utilized, but on UMLS.

Let *N* be an adjacency matrix for the UMLS graph with all relations in MRREL. The random walker starts in any of the concepts included in the sentence, and randomly follows 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

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

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

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 2: 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.*"

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 $\vec{u}^{(0)}$. Having 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 2.

$$\vec{u}^{(t)} = (1-\beta)N\vec{u}^{(t-1)} + \beta\vec{u}^{(0)} \tag{2}$$

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.

Semantic Similarities at the Sentence Level For comparing pairs of semantic signatures at the sentence level, we use Weighted Overlap (WO) algorithm proposed by (Pilehvar and Navigli, 2015). This algorithm first sorts the two signatures according to their values and then harmonically weights the overlaps between them. Using the knowledge source N (i.e. WordNet or UMLS), WO calculates the semantic similarity (Sim_N) of two sentence signatures S_{N1} and S_{N2} as:

$$Sim_N(S_{N1}, S_{N2}) = \frac{\sum_{h \in H} (r_h(S_{N1}) + r_h(S_{N2}))^{-1}}{\sum_{i=1}^{|H|} (2i)^{-1}} \tag{3}$$

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 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 4.

$$Sim_{final}(S_1, S_2) = \mu \times Sim_U(S_{U1}, S_{U2}) + (1-\mu) \times Sim_W(S_{W1}, S_{W2}) \tag{4}$$

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 the WordNet is also shown by $Sim_W(S_{W1}, S_{W2})$. The scaling factor μ was optimized on development data in our experiments and set to 0.6 to reach the best result (Section 5).

4.2 Constructing Similarity Graph

To filter out less query relevant information, sentences are modeled as a Similarity Graph - a weighted undirected graph on which each node represents a sentence and the edge weight carries the similarity

of two sentences (ShafieiBavani et al., 2016b). 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 4.1). 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. 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 employ a combination model (Chali et al., 2011):

$$C(S_i|Q) = \gamma \times \frac{Sim_{final}(S_i, Q)}{\sum_{S_j \in A} Sim_{final}(S_j, Q)} + (1 - \gamma) \times \sum_{S_k \in A} \frac{Sim_{final}(S_i, S_k)}{\sum_{S_j \in A} Sim_{final}(S_j, S_k)} \times C(S_k|Q) \tag{5}$$

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 (Chali et al., 2011) 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 and 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 results in a subgraph comprising a set of the most query-relevant sentences to be clustered in the next step.

4.3 Clustering Relevant Sentences

In this step, we use Chinese Whispers (CW) which is a graph-based clustering algorithm proposed by (Biemann, 2006). CW is a 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, 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. First, a distinct class is assigned to each node , and a clustering C containing the singleton clusters c_i is created. Then, a series of iterations is performed to merge the clusters. Specifically, at each iteration the algorithm analyzes 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 (Equation 6).

$$class(s_j) = \underset{c}{argmax} \sum_{\substack{\{s_i, s_j\} \in E(G) \\ s.t. class(s_i) = c}} Sim(s_i, s_j) \tag{6}$$

As soon as an iteration produces no change in the clustering, the algorithm stops and outputs the final clustering. 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.

4.4 Word Graph-based Summarization of EBM

For each obtained cluster, we build a word graph by iteratively adding sentences to it. This graph is an ordered pair $G = (V, E)$ comprising of a set of vertices (nodes) V, together with a set of directed edges E, 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 word 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 3, for the last group, we use 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 (Filippova, 2010): 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 1 (a)). Edge weights are calculated using the weighting function defined in Equation 7 (Filippova, 2010).

$$W(e_{i,j}) = \frac{(freq(i) + freq(j))/\sum_{s \in S} diff(s,i,j)^{-1}}{freq(i) \times freq(j)} \tag{7}$$

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 if available. 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, 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 (Figure 1 (b)).

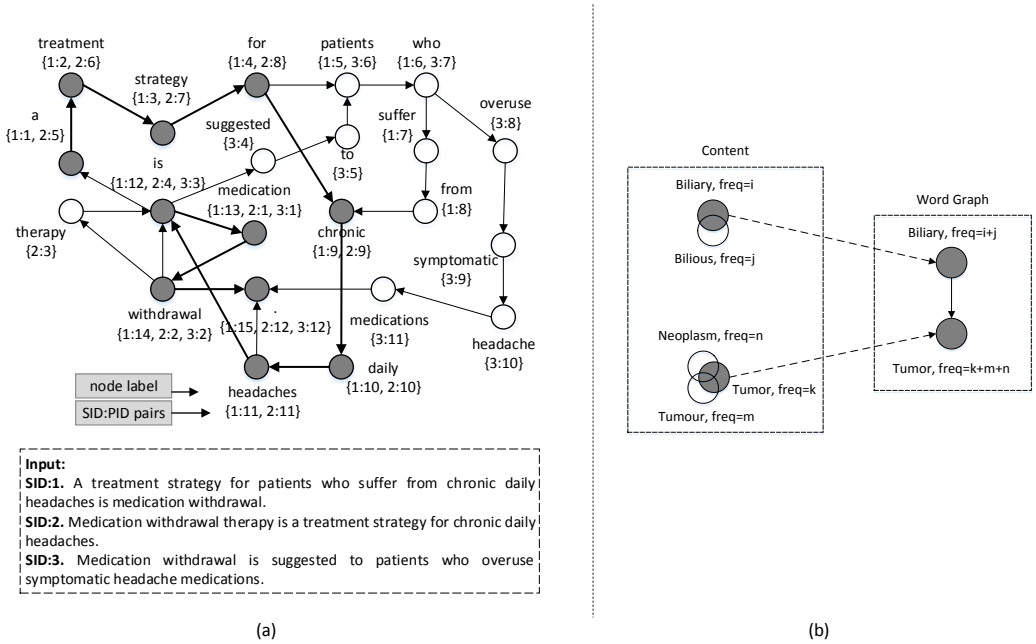

Figure 1: (a) An example of the Constructed Word Graph. Thick edges indicate salient paths. (b) An example of Biomedical Synonym Mapping

Ensuring Information Richness To re-rank the summary candidates based on the information richness, important key-phrases have been exploited using the TextRank algorithm (Mihalcea and Tarau, 2004). 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 summary candidate c is given by Equation 8.

$$Score_{Key}(c) = \frac{\sum_{i,j \in path(c)} W(e_{i,j})}{length(c) \times \sum_{k \in c} (\frac{\sum_{w \in k} TextRank(W)}{length(k)+1})} \quad (8)$$

The heuristic algorithm discussed in (Boudin and Morin, 2013) is then used to find the k-shortest paths ($k = 50$ throughout our experiments) from start to end node in the graph. 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 richest summary sentence.

Ensuring Syntactic Structure Since our 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 (ShafieiBavani et al., 2016a). 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-richest sentences. It estimates the probability of string of m POS tags by $p(t_1^m) \propto \prod_{i-1}^{m} p(t_i|t_{i-n+1}^{i-1})$ (Monz, 2011), where *n* is 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 make use of Stanford POS tagger to annotate a large part (∼100 M-words) of the BioMed Central full-text corpus for text mining research[11]. Then, we remove all words from the pairs of words/POS in the POS annotated corpus. Finally, the SRILM toolkit (Stolcke and others, 2002) is employed to collect n-gram statistics. The candidate sentences also need to be annotated with POS tags, and the score of each summary is estimated by the 7-gram language modeling, based on its sequence of POS tags. 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-richest sentences is computed by Equation 9.

$$Score_{LM}(c) = 10^{\frac{\log prob(c)}{|word|}} \quad (9)$$

where $prob(c)$ is the probability of summary candidate 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 8, and the score of POS-LM into the range [0, 1]. The score of each summary is finally given by Equation 10.

$$Score_{final}(c) = \eta \times Score_{Key}(c) + (1 - \eta) \times Score_{LM}(c) \quad (10)$$

The scaling factor η was optimized on development data in our experiments and set to 0.4 (Section 5). Hence, the most grammatical candidate among the candidates that 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 (Plaza et al., 2011). 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.

5 Experiments

In our work, the generated summaries are assessed automatically through version 2.0 [12] of ROUGE (Lin, 2004) over the released EBM corpus by (Mollá et al., 2015). ROUGE measures the summary quality by counting the overlapping units between system-generated summaries and human-written reference/gold summaries. We 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.

To investigate the effectiveness of our approach, we compare our summarizer with *FastSum* (Schilder and Kondadadi, 2008), and a research prototype *LexRank* (Erkan and Radev, 2004). FastSum is a fast query-focused multi-document summarization system based only on word frequency features of topics,

[11] http://old.biomedcentral.com/about/datamining
[12] http://kavita-ganesan.com/content/rouge-2.0

documents, and clusters. Each sentence is ranked based on a linear function of scores using a variety of frequency measures. A support vector machine regression is also used to learn weights of the features. Comparing our approach with FastSum would let us evaluate the superiority of our approach over the word frequency-based approaches on the task of query-focused multi-document summarization. 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 outperformed both centroid-based methods and other systems participating in DUC in most of the cases (Erkan and Radev, 2004). Comparison with LexRank will allow us to evaluate whether semantic information provides benefits over merely lexical information in graph-based summarization approaches. Table 3 shows an example of a summary generated by human (Gold), our proposed approach (Proposed), and 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 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 3: An example of Gold summary, Proposed summary, and LexRank summary

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. In details, 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 of 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 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 (Demner-Fushman and Lin, 2007). In our work, this baseline generates summaries by selecting three sentences at random from the *last part*.

The average performance of the baseline systems and the proposed approach in terms of ROUGE scores are provided in Table 4.

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	**0.3985**	**0.2450**	**0.2259**

Table 4: Average scores by ROUGE metrics over the EBM corpus

The statistics point out the effectiveness of our summarizer over the compared systems on all evaluation metrics. Besides, 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.

Standard Deviation of ROUGE Scores Since Table 4 demonstrates 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 the proposed approach are shown in Table 5.

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

Table 5: Standard deviation of ROUGE scores for the summaries generated by the proposed approach

Exploring Scaling Factors In our work, two free parameters are defined: (1) μ for measuring semantic similarities using WordNet and UMLS; (2) η for final re-ranking score of each generated summary sentence. We randomly selected 30% of the EBM corpus as our development set to tune these parameters. Figure 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 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 the 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.

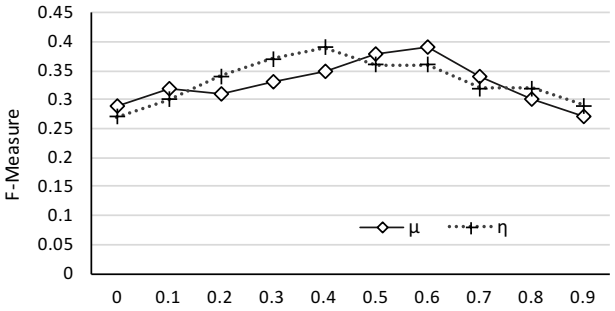

Figure 2: Exploring scaling factors μ and η on the development set

6 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; (4) summarizing 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 approach outperforms the two competitive systems. It has also been found that the last part of each abstract is more likely to be included in the summary. 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) (Plaza et al., 2011). 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.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions.

References

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

Chris Biemann. 2006. 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.

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

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

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

Yllias Chali, Sadid A Hasan, and Shafiq R Joty. 2011. 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.

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

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

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

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Katja Filippova. 2010. 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.

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

Evidence-Based Medicine Working Group and Others. 1992. Evidence-based medicine. a new approach to teaching the practice of medicine. *The Journal of the American Medical Association*, 268(17):2420.

Allan Hanbury. 2012. 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.

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

William R Hersh, M Katherine Crabtree, David H Hickam, Lynetta Sacherek, Charles P Friedman, Patricia Tidmarsh, Craig Mosbaek, and Dale Kraemer. 2002. 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.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8.

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

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

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

Christof Monz. 2011. 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.

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

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

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

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

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

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

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

Frank Schilder and Ravikumar Kondadadi. 2008. 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.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2016a. An efficient approach for multi-sentence compression. In *JMLR: Workshop and Conference Proceedings*.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2016b. A query-based summarization service from multiple news sources. In *Services Computing (SCC), 2016 IEEE International Conference on*, pages 42–49. IEEE.

Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M Kashani, Anoop Sarkar, and Fred Popowich. 2007. Question answering summarization of multiple biomedical documents. In *Advances in Artificial Intelligence*, pages 284–295. Springer.

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

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

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