Improving Summarization of Biomedical Documents using Word Sense Disambiguation

Laura Plaza[†] lplazam@fdi.ucm.es
Mark Stevenson[*] m.stevenson@dcs.shef.ac.uk
Alberto Díaz[†] albertodiaz@fdi.ucm.es
[†] Universidad Complutense de Madrid, C/Prof. José García Santesmases, 28040 Madrid, Spain
[*] University of Sheffield, Regent Court, 211 Portobello St., Sheffield, S1 4DP, UK

Abstract

We describe a concept-based summarization system for biomedical documents and show that its performance can be improved using Word Sense Disambiguation. The system represents the documents as graphs formed from concepts and relations from the UMLS. A degree-based clustering algorithm is applied to these graphs to discover different themes or topics within the document. To create the graphs, the MetaMap program is used to map the text onto concepts in the UMLS Metathesaurus. This paper shows that applying a graph-based Word Sense Disambiguation algorithm to the output of MetaMap improves the quality of the summaries that are generated.

1 Introduction

Extractive text summarization can be defined as the process of determining salient sentences in a text. These sentences are expected to condense the relevant information regarding the main topic covered in the text. Automatic summarization of biomedical texts may benefit both health-care services and biomedical research (Reeve et al., 2007; Hunter and Cohen, 2006). Providing physicians with summaries of their patient records can help to reduce the diagnosis time. Researchers can use summaries to quickly determine whether a document is of interest without having to read it all.

Summarization systems usually work with a representation of the document consisting of information that can be directly extracted from the document itself (Erkan and Radev, 2004; Mihalcea and Tarau, 2004). However, recent studies have demonstrated the benefit of summarization based on richer representations that make use of external knowledge sources (Plaza et al., 2008; Fiszman et al., 2004). These approaches can represent semantic associations between the words and terms in the document (i.e. synonymy, hypernymy, homonymy or co-occurrence) and use this information to improve the quality of the summaries. In the biomedical domain the Unified Medical Language System (UMLS) (Nelson et al., 2002) has proved to be a useful knowledge source for summarization (Fiszman et al., 2004; Reeve et al., 2007; Plaza et al., 2008). In order to access the information contained in the UMLS, the vocabulary of the document being summarized has to be mapped onto it. However, ambiguity is common in biomedical documents (Weeber et al., 2001). For example, the string *"cold"* is associated with seven possible meanings in the UMLS Metathesuarus including *"common cold"*, *"cold sensation"* , *"cold temperature"* and *"Chronic Obstructive Airway Disease"*. The majority of summarization systems in the biomedical domain rely on MetaMap (Aronson, 2001) to map the text onto concepts from the UMLS Metathesaurus (Fiszman et al., 2004; Reeve et al., 2007). However, MetaMap frequently fails to identify a unique mapping and, as a result, various concepts with the same score are returned. For instance, for the phrase "*tissues are often cold*" MetaMap returns three equally scored concepts for the word ''cold": *"common cold"*, *"cold sensation"* and *"cold temperature"*.

The purpose of this paper is to study the effect of lexical ambiguity in the knowledge source on semantic approaches to biomedical summarization. To this end, the paper describes a concept-based summarization system for biomedical documents that uses the UMLS as an external knowledge source. To address the word ambiguity problem, we have adapted an existing WSD system (Agirre and Soroa, 2009) to assign concepts from the UMLS. The system is applied to the summarization of 150 biomedical scientific articles from the BioMed Central corpus and it is found that

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 55–63,
Uppsala, Sweden, 15 July 2010.

WSD improves the quality of the summaries. This paper is, to our knowledge, the first to apply WSD to the summarization of biomedical documents and also demonstrates that this leads to an improvement in performance.

The next section describes related work on summarization and WSD. Section 3 introduces the UMLS resources used in the WSD and summarization systems. Section 4 describes our concept-based summarization algorithm. Section 5 presents a graph-based WSD algorithm which has been adapted to assign concepts from the UMLS. Section 6 describes the experiments carried out to evaluate the impact of WSD and discusses the results. The final section provides concluding remarks and suggests future lines of work.

2 Related work

Summarization has been an active area within NLP research since the 1950s and a variety of approaches have been proposed (Mani, 2001; Afantenos et al., 2005). Our focus is on graph-based summarization methods. Graph-based approaches typically represent the document as a graph, where the nodes represent text units (i.e. words, sentences or paragraphs), and the links represent cohesion relations or similarity measures between these units. The best-known work in the area is LexRank (Erkan and Radev, 2004). It assumes a fully connected and undirected graph, where each node corresponds to a sentence, represented by its *TF-IDF* vector, and the edges are labeled with the cosine similarity between the sentences. Mihalcea and Tarau (2004) present a similar method where the similarity among sentences is measured in terms of word overlaps.

However, methods based on term frequencies and syntactic representations do not exploit the semantic relations among the words in the text (i.e. synonymy, homonymy or co-occurrence). They cannot realize, for instance, that the phrases *myocardial infarction* and *heart attack* refer to the same concepts, or that *pneumococcal pneumonia* and *mycoplasma pneumonia* are two similar diseases that differ in the type of bacteria that causes them. This problem can be partially solved by dealing with concepts and semantic relations from domain-specific resources, rather than terms and lexical or syntactic relations. Consequently, some recent approaches have adapted existing methods to represent the document at a conceptual level. In particular, in the biomedical domain Reeve et al. (2007) adapt the lexical chaining approach (Barzilay and Elhadad, 1997) to work with UMLS concepts, using the MetaMap Transfer Tool to annotate these concepts. Yoo et al. (2007) represent a corpus of documents as a graph, where the nodes are the MeSH descriptors found in the corpus, and the edges represent hypernymy and co-occurrence relations between them. They cluster the MeSH concepts in the corpus to identify sets of documents dealing with the same topic and then generate a summary from each document cluster.

Word sense disambiguation attempts to solve lexical ambiguities by identifying the correct meaning of a word based on its context. Supervised approaches have been shown to perform better than unsupervised ones (Agirre and Edmonds, 2006) but need large amounts of manually-tagged data, which are often unavailable or impractical to create. Knowledge-based approaches are a good alternative that do not require manually-tagged data.

Graph-based methods have recently been shown to be an effective approach for knowledge-based WSD. They typically build a graph for the text in which the nodes represent all possible senses of the words and the edges represent different kinds of relations between them (e.g. lexico-semantic, co-occurrence). Some algorithm for analyzing these graphs is then applied from which a ranking of the senses of each word in the context is obtained and the highest-ranking one is chosen (Mihalcea and Tarau, 2004; Navigli and Velardi, 2005; Agirre and Soroa, 2009). These methods find globally optimal solutions and are suitable for disambiguating all words in a text.

One such method is Personalized PageRank (Agirre and Soroa, 2009) which makes use of the PageRank algorithm used by internet search engines (Brin and Page, 1998). PageRank assigns weight to each node in a graph by analyzing its structure and prefers ones that are linked to by other nodes that are highly weighted. Agirre and Soroa (2009) used WordNet as the lexical knowledge base and creates graphs using the entire WordNet hierarchy. The ambiguous words in the document are added as nodes to this graph and directed links are created from them to each of their possible meanings. These nodes are assigned weight in the graph and the PageRank algorithm is

applied to distribute this information through the graph. The meaning of each word with the highest weight is chosen. We refer to this approach as ppr. It is efficient since it allows all ambiguous words in a document to be disambiguated simultaneously using the whole lexical knowledge base, but can be misled when two of the possible senses for an ambiguous word are related to each other in WordNet since the PageRank algorithm assigns weight to these senses rather than transferring it to related words. Agirre and Soroa (2009) also describe a variant of the approach, referred to as "word to word" (ppr_w2w), in which a separate graph is created for each ambiguous word. In these graphs no weight is assigned to the word being disambiguated so that all of the information used to assign weights to the possible senses of the word is obtained from the other words in the document. The ppr_w2w is more accurate but less efficient due to the number of graphs that have to be created and analyzed. Agirre and Soroa (2009) show that the Personalized PageRank approach performs well in comparison to other knowledge-based approaches to WSD and report an accuracy of around 58% on standard evaluation data sets.

3 UMLS

The Unified Medical Language System (UMLS) (Humphreys et al., 1998) is a collection of controlled vocabularies related to biomedicine and contains a wide range of information that can be used for Natural Language Processing. The UMLS comprises of three parts: the Specialist Lexicon, the Semantic Network and the Metathesaurus.

The **Metathesaurus** forms the backbone of the UMLS and is created by unifying over 100 controlled vocabularies and classification systems. It is organized around concepts, each of which represents a meaning and is assigned a Concept Unique Identifier (CUI). For example, the following CUIs are all associated with the term "cold": C0009443 'Common Cold', C0009264 'Cold Temperature' and C0234192 'Cold Sensation'.

The MRREL table in the Metathesaurus lists relations between CUIs found in the various sources that are used to form the Metathesaurus. This table lists a range of different types of relations, including CHD ("child"), PAR ("parent"), QB ("can be qualified by"), RQ ("related and possibly synonymous") and RO ("other related"). For example, the MRREL table states that C0009443 'Common Cold' and C0027442 'Nasopharynx' are connected via the RO relation.

The MRHIER table in the Metathesaurus lists the hierarchies in which each CUI appears, and presents the whole path to the top or root of each hierarchy for the CUI. For example, the MRHIER table states that C0035243 'Respiratory Tract Infections' is a parent of C0009443 'Common Cold'.

The **Semantic Network** consists of a set of categories (or semantic types) that provides a consistent categorization of the concepts in the Metathesaurus, along with a set of relationships (or semantic relations) that exist between the semantic types. For example, the CUI C0009443 'Common Cold' is classified in the semantic type 'Disease or Syndrome'.

The SRSTR table in the Semantic Network describes the structure of the network. This table lists a range of different relations between semantic types, including hierarchical relations (is_a) and non hierarchical relations (e.g. result_of, associated_with and co-occurs_with). For example, the semantic types 'Disease or Syndrome' and 'Pathologic Function' are connected via the is_a relation in this table.

4 Summarization system

The method presented in this paper consists of 4 main steps: (1) concept identification, (2) document representation, (3) concept clustering and topic recognition, and (4) sentence selection. Each step is discussed in detail in the following subsections.

4.1 Concept identification

The first stage of our process is to map the document to concepts from the UMLS Metathesaurus and semantic types from the UMLS Semantic Network.

We first run the MetaMap program over the text in the body section of the document[1] MetaMap (Aronson, 2001) identifies all the phrases that could be mapped onto a UMLS CUI, retrieves and scores all possible CUI mappings for each phrase, and returns all the candidates along with

[1]We do not make use of the disambiguation algorithm provided by MetaMap, which is invoked using the -y flag (Aronson, 2006), since our aim is to compare the effect of WSD on the performance of our summarization system rather than comparing WSD algorithms.

their score. The semantic type for each concept mapping is also returned. Table 1 shows this mapping for the phrase *tissues are often cold*. This example shows that MetaMap returns a single CUI for two words (*tissues* and *often*) but also returns three equally scored CUIs for *cold* (C0234192, C0009443 and C0009264). Section 5 describes how concepts are selected when MetaMap is unable to return a single CUI for a word.

Phrase: "Tissues"
Meta Mapping (1000)
1000 C0040300:Tissues (Body tissue)
Phrase: "are"
Phrase: "often cold"
MetaMapping (888)
694 C0332183:Often (Frequent)
861 C0234192:Cold (Cold Sensation)
MetaMapping (888)
694 C0332183:Often (Frequent)
861 C0009443:Cold (Common Cold)
MetaMapping (888)
694 C0332183:Often (Frequent)
861 C0009264:Cold (cold temperature)

Table 1: An example of MetaMap mapping for the phrase *Tissues are often cold*

UMLS concepts belonging to very general semantic types are discarded, since they have been found to be excessively broad or unrelated to the main topic of the document. These types are *Quantitative Concept*, *Qualitative Concept*, *Temporal Concept*, *Functional Concept*, *Idea or Concept*, *Intellectual Product*, *Mental Process*, *Spatial Concept* and *Language*. Therefore, the concept C0332183 'Often' in the previous example, which belongs to the semantic type *Temporal Concept*, is discarded.

4.2 Document representation

The next step is to construct a graph-based representation of the document. To this end, we first extend the disambiguated UMLS concepts with their complete hierarchy of hypernyms and merge the hierarchies of all the concepts in the same sentence to construct a graph representing it. The two upper levels of these hierarchies are removed, since they represent concepts with excessively broad meanings and may introduce noise to later processing.

Next, all the sentence graphs are merged into a single document graph. This graph is extended with more semantic relations to obtain a more complete representation of the document. Various types of information from the UMLS can be used to extend the graph. We experimented using different sets of relations and finally used the *hypernymy* and *other related* relations between concepts from the Metathesaurus, and the *associated with* relation between semantic types from the Semantic Network. Hypernyms are extracted from the `MRHIER` table, `RO` ("other related") relations are extracted from the `MRREL` table, and `associated_with` relations are extracted from the `SRSTR` table (see Section 3). Finally, each edge is assigned a weight in [0, 1]. This weight is calculated as the ratio between the relative positions in their corresponding hierarchies of the concepts linked by the edge.

Figure 1 shows an example graph for a simplified document consisting of the two sentences below. Continuous lines represent *hypernymy* relations, dashed lines represent *other related* relations and dotted lines represent *associated with* relations.

1. The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat hypertension.

2. The trial was carried out in two groups: the first group taking doxazosin, and the second group taking chlorthalidone.

4.3 Concept clustering and topic recognition

Our next step consists of clustering the UMLS concepts in the document graph using a *degree-based clustering* method (Erkan and Radev, 2004). The aim is to construct sets of concepts strongly related in meaning, based on the assumption that each of these sets represents a different topic in the document.

We assume that the document graph is an instance of a *scale-free network* (Barabasi and Albert, 1999). A scale-free network is a complex network that (among other characteristics) presents a particular type of node which are highly connected to other nodes in the network, while the remaining nodes are quite unconnected. These highest-degree nodes are often called *hubs*. This scale-free power-law distribution has been empirically observed in many large networks, including linguistic and semantic ones.

To discover these prominent or hub nodes, we compute the *salience* or prestige of each vertex

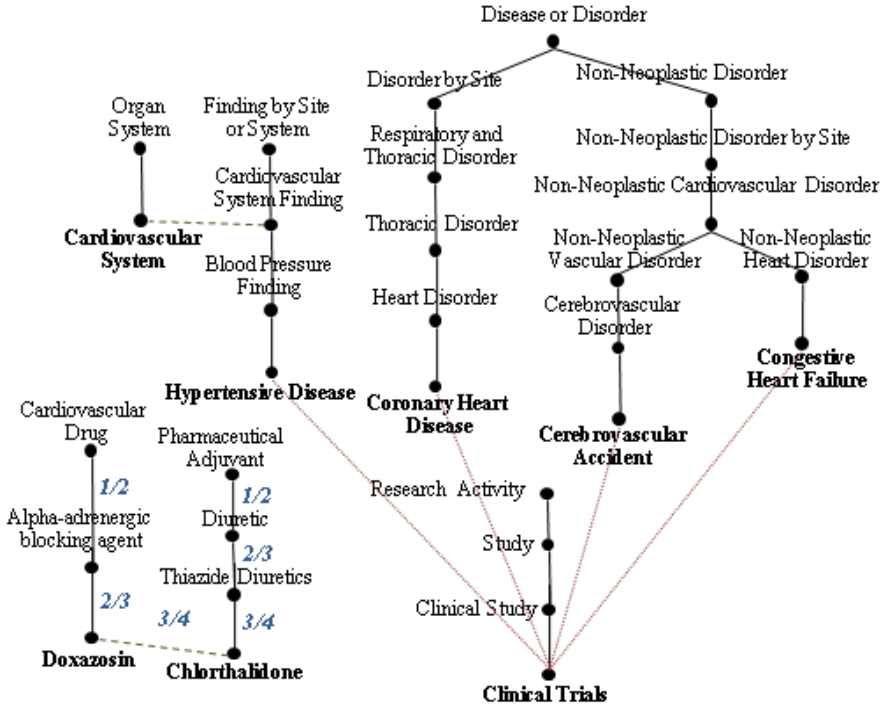

Figure 1: Example of a simplified document graph

in the graph (Yoo et al., 2007), as shown in (1). Whenever an edge from v_i to v_j exists, a vote from node *i* to node *j* is added with the strength of this vote depending on the weight of the edge. This ranks the nodes according to their structural importance in the graph.

$$salience(v_i) = \sum_{\forall e_j | \exists v_k \wedge e_j connect(v_i, v_k)} weight(e_j) \quad (1)$$

The *n* vertices with a highest salience are named *Hub Vertices*. The clustering algorithm first groups the hub vertices into *Hub Vertices Sets (HVS)*. These can be seen as set of concepts strongly related in meaning, and will represent the centroids of the clusters. To construct these HVS, the clustering algorithm first searches, iteratively and for each hub vertex, the hub vertex most connected to it, and merges them into a single HVS. Second, the algorithm checks, for every pair of HVS, if their internal connectivity is lower than the connectivity between them. If so, both HVS are merged. The remaining vertices (i.e. those not included in the HVS) are iteratively assigned to the cluster to which they are more connected. This connectivity is computed as the sum of the weights of the edges that connect the target vertex to the other vertices in the cluster.

4.4 Sentence selection

The last step of the summarization process consists of computing the similarity between all sentences in the document and each of the clusters, and selecting the sentences for the summary based on these similarities. To compute the similarity between a sentence graph and a cluster, we use a non-democratic vote mechanism (Yoo et al., 2007), so that each vertex of a sentence assigns a vote to a cluster if the vertex belongs to its HVS, half a vote if the vertex belongs to it but not to its HVS, and no votes otherwise. Finally, the similarity between the sentence and the cluster is computed as the sum of the votes assigned by all the vertices in the sentence to the cluster, as expressed in (2).

$$similarity(C_i, S_j) = \sum_{v_k | v_k \in S_j} w_{k,j} \quad (2)$$

$$\text{where} \begin{cases} w_{k,j}=0 \, if \, v_k \notin C_i \\ w_{k,j}=1 \, if \, v_k \in HVS(C_i) \\ w_{k,j}=0.5 \, if \, v_k \notin HVS(C_i) \end{cases}$$

Finally, we select the sentences for the summary based on the similarity between them and the clusters as defined above. In previous work (blind reference), we experimented with different heuristics for sentence selection. In this paper, we just present the one that reported the best results. For each sentence, we compute a single score, as

the sum of its similarity to each cluster adjusted to the cluster's size (expression 3). Then, the *N* sentences with higher scores are selected for the summary.

$$Score(S_j) = \sum_{C_i} \frac{similarity(C_i, S_j)}{|C_i|} \quad (3)$$

In addition to semantic-graph similarity (*SemGr*) we have also tested two further features for computing the salience of sentences: sentence location (*Location*) and similarity with the title section (*Title*). The sentence location feature assigns higher scores to the sentences close to the beginning and the end of the document, while the similarity with the title feature assigns higher scores as the proportion of common concepts between the title and the target sentence is increased. Despite their simplicity, these are well accepted summarization heuristics that are commonly used (Bawakid and Oussalah, 2008; Bossard et al., 2008).

The final selection of the sentences for the summary is based on the weighted sum of these feature values, as stated in (4). The values for the parameters λ, θ and χ have been empirically set to 0.8, 0.1, and 0.1 respectively.

$$Score(S_j) = \lambda \times SemGr(S_j) + \theta \times Location(S_j) + \chi \times Title(S_j) \quad (4)$$

5 WSD for concept identification

Since our summarization system is based on the UMLS it is important to be able to accurately map the documents onto CUIs. The example in Section 4.1 shows that MetaMap does not always select a single CUI and it is therefore necessary to have some method for choosing between the ones that are returned. Summarization systems typically take the first mapping as returned by MetaMap, and no attempt is made to solve this ambiguity (Plaza et al., 2008). This paper reports an alternative approach that uses a WSD algorithm that makes use of the entire UMLS Metathesaurus.

The Personalized PageRank algorithm (see Section 2) was adapted to use the UMLS Metathesaurus and used to select a CUI from the MetaMap output[2]. The UMLS is converted into a graph in which the CUIs are the nodes and the edges are derived from the `MRREL` table. All possible relations in this table are included. The output from MetaMap is used to provide the list of possible CUIs for each term in a document and these are passed to the disambiguation algorithm. We use both the standard (`ppr`) and "word to word" (`ppr_w2w`) variants of the Personalized PageRank approach.

It is difficult to evaluate how well the Personalized PageRank approach performs when used in this way due to a lack of suitable data. The NLM-WSD corpus (Weeber et al., 2001) contains manually labeled examples of ambiguous terms in biomedical text but only provides examples for 50 terms that were specifically chosen because of their ambiguity. To evaluate an approach such as Personalized PageRank we require documents in which the sense of every ambiguous word has been identified. Unfortunately no such resource is available and creating one would be prohibitively expensive. However, our main interest is in whether WSD can be used to improve the summaries generated by our system rather than its own performance and, consequently, decided to evaluate the WSD by comparing the output of the summarization system with and without WSD.

6 Experiments

6.1 Setup

The ROUGE metrics (Lin, 2004) are used to evaluate the system. ROUGE compares automatically generated summaries (called *peers*) against human-created summaries (called *models*), and calculates a set of measures to estimate the content quality of the summaries. Results are reported for the ROUGE-1 (**R-1**), ROUGE-2 (**R-2**), ROUGE-SU4 (**R-SU**) and ROUGE-W (**R-W**) metrics. ROUGE-N (e.g. ROUGE-1 and ROUGE-2) evaluates n-gram co-occurrences among the peer and models summaries, where N stands for the length of the n-grams. ROUGE-SU4 allows bi-gram to have intervening word gaps no longer than four words. Finally, ROUGE-W computes the union of the longest common subsequences between the candidate and the reference summaries taking into account the presence of consecutive matches.

To the authors' knowledge, no specific corpus for biomedical summarization exists. To evaluate our approach we use a collection of 150 documents randomly selected from the BioMed Cen-

[2]We use a publicly available implementation of the Personalized Page Rank algorithm (`http://ixa2.si.ehu.es/ukb/`) for the experiments described here.

tral corpus[3] for text mining research. This collection is large enough to ensure significant results in the ROUGE evaluation (Lin, 2004) and allows us to work with the ppr_w2w disambiguation software, which is quite time consuming. We generate automatic summaries by selecting sentences until the summary reaches a length of the 30% over the original document size. The abstract of the papers (i.e. the authors' summaries) are removed from the documents and used as model summaries.

A separate development set was used to determine the optimal values for the parameters involved in the algorithm. This set consists of 10 documents from the BioMed Central corpus. The model summaries for these documents were manually created by medical students by selecting between 20-30% of the sentences within the paper. The parameters to be estimated include the percentage of vertices considered as hub vertices by the clustering method (see Section 4.3) and the combination of summarization features used to sentence selection (see Section 4.4). As a result, the percentage of hub vertices was set to 15%, and no additional summarization features (apart from the semantic-graph similarity) were used.

Two baselines were also implemented. The first, *lead baseline*, generate summaries by selecting the first *n* sentences from each document. The second, *random baseline*, randomly selects *n* sentences from the document. The *n* parameter is based on the desired compression rate (i.e. 30% of the document size).

6.2 Results

Various summarizers were created and evaluated. First, we generated summaries using our method without performing word sense disambiguation (SemGr), but selecting the first CUI returned by MetaMap. Second, we repeated these experiments using the Personalized Page Rank disambiguation algorithm (ppr) to disambiguate the CUIs returned by MetaMap (SemGr + ppr). Finally, we use the "word to word" variant of the Personalized Page Rank algorithm (ppr_w2w) to perform the disambiguation (SemGr + ppr_w2w).

Table 2 shows ROUGE scores for the different configurations of our system together with the two baselines. All configurations significantly outperform both baselines (Wilcoxon Signed Ranks Test, $p < 0.01$).

[3] http://www.biomedcentral.com/info/about/datamining/

Summarizer	R-1	R-2	R-W	R-SU
random	.5089	.1879	.1473	.2349
lead	.6483	.2566	.1621	.2646
SemGr	.7504	.3283	.1915	.3117
SemGr+ppr	**.7737**	**.3419**	.1937	.3178
SemGr+ppr_w2w	**.7804**	**.3530**	**.1966**	**.3262**

Table 2: ROUGE scores for two baselines and SemGr (with and without WSD). Significant differences among the three versions of SemGr are indicated in bold font.

The use of WSD improves the average ROUGE score for the summarizer. The "standard" (i.e. ppr) version of the WSD algorithm significantly improves ROUGE-1 and ROUGE-2 metrics (Wilcoxon Signed Ranks Test, $p < 0.01$), compared with no WSD (i.e. SemGr). The "word to word" variant (ppr_w2w) significantly improves all ROUGE metrics. Performance using the "word to word" variant is also higher than standard ppr in all ROUGE scores.

These results demonstrate that employing a state of the art WSD algorithm that has been adapted to use the UMLS Metathesaurus improves the quality of the summaries generated by a summarization system. To our knowledge this is the first result to demonstrate that WSD can improve summarization systems. However, this improvement is less than expected and this is probably due to errors made by the WSD system. The Personalized PageRank algorithms (ppr and ppr_w2w) have been reported to correctly disambiguate around 58% of words in general text (see Section 2) and, although we were unable to quantify their performance when adapted for the biomedical domain (see Section 5), it is highly likely that they will still make errors. However, the WSD performance they do achieve is good enough to improve the summarization process.

6.3 Analysis

The results presented above demonstrate that using WSD improves the performance of our summarizer. The reason seems to be that, since the accuracy in the concept identification step increases, the document graph built in the following steps is a better approximation of the structure of the document, both in terms of concepts and relations. As a result, the clustering method succeeds in finding the topics covered in the document, and the information in the sentences selected for the summary

is closer to that presented in the model summaries.

We have observed that the clustering method usually produces one big cluster along with a variable number of small clusters. As a consequence, though the heuristic for sentence selection was designed to select sentences from all the clusters in the document, the fact is that most of the sentences are extracted from this single large cluster. This allows our system to identify sentences that cover the main topic of the document, while it occasionally fails to extract other "satellite" information.

We have also observed that the ROUGE scores differ considerably from one document to others. To understand the reasons of these differences we examined the two documents with the highest and lowest ROUGE scores respectively. The best case is one of the largest document in the corpus, while the worst case is one of the shortest (6 versus 3 pages). This was expected, since according to our hypothesis that the document graph is an instance of a scale-free network (see Section 4.3), the summarization algorithm works better with larger documents. Both documents also differ in their underlying subject matter. The best case concerns the reactions of some kind of proteins over the brain synaptic membranes; while the worst case regards the use of pattern matching for database searching. We have verified that UMLS covers the vocabulary contained in the first document better than in the second one. We have also observed that the use in the abstract of synonyms of terms presented in the document body is quite frequent. In particular the worst case document uses different terms in the abstract and the body, for example "pattern matching" and "string searching". Since the ROUGE metrics rely on evaluating summaries based on the number of strings they have in common with the model summaries the system's output is unreasonably penalised.

Another problem is related to the use of acronyms and abbreviations. Most papers in the corpus do not include an *Abbreviations* section but define them *ad hoc* in the document body. These contracted forms are usually non-standard and do not exist in the UMLS Metathesaurus. This seriously affects the performance of both the disambiguation and the summarization algorithms, especially considering that it has been observed that the terms (or phrases) represented in an abbreviated form frequently correspond to central concepts in the document. For example, in a paper from the corpus that presents an analysis tool for simple sequence repeat tracts in DNA, only the first occurrence of 'simple sequence repeat' is presented in its expanded form. In the remaining of the document, this phrase is named by its acronym 'SSR'. The same occurs in a paper that investigates the developmental expression of survivin during embryonic submandibular salivary gland development, where 'embryonic submandibular gland' is always referred as 'SMG'.

7 Conclusion and future work

In this paper we propose a graph-based approach to biomedical summarization. Our algorithm represents the document as a semantic graph, where the nodes are concepts from the UMLS Metathesaurus and the links are different kinds of semantic relations between them. This produces a richer representation than the one provided by traditional models based on terms.

This approach relies on accurate mapping of the document being summarized into the concepts in the UMLS Metathesaurus. Three methods for doing this were compared and evaluated. The first was to select the first mapping generated by MetaMap while the other two used a state of the art WSD algorithm. This WSD algorithm was adapted for the biomedical domain by using the UMLS Metathesaurus as a knowledge based and MetaMap as a pre-processor to identify the possible CUIs for each term. Results show that the system performs better when WSD is used.

In future work we plan to make use of the different types of information within the UMLS to create different configurations of the Personalized PageRank WSD algorithm and explore their effect on the summarization system (i.e. considering different UMLS relations and assigning different weights to different relations). It would also be interesting to test the system with other disambiguation algorithms and use a state of the art algorithm for identifying and expanding acronyms and abbreviations.

Acknowledgments

This research is funded by the Spanish Government through the FPU program and the projects TIN2009-14659-C03-01 and TSI 020312-2009-44. Mark Stevenson acknowledges the support of the Engineering and Physical Sciences Research Council (grant EP/D069548/1).

References

S.D. Afantenos, V. Karkaletsis, and P. Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

E. Agirre and P. Edmonds, editors, 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

E. Agirre and A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of EACL-09*, pages 33–41, Athens, Greece.

A. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, pages 17–21.

A. Aronson. 2006. MetaMap: Mapping text to the UMLS Metathesaurus. Technical report, U.S. National Library of Medicine.

A.L. Barabasi and R. Albert. 1999. Emergence of scaling in random networks. *Science*, 268:509–512.

R. Barzilay and M. Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

A. Bawakid and M. Oussalah. 2008. A semantic summarization system: University of Birmingham at TAC 2008. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

A. Bossard, M. Gnreux, and T. Poibeau. 2008. Description of the LIPN systems at TAC 2008: summarizing information and opinions. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:1–7.

G. Erkan and D. R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22:457–479.

M. Fiszman, T. C. Rindflesch, and H. Kilicoglu. 2004. Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 76–83.

L. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. 1998. The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*, 1(5):1–11.

L. Hunter and K. B. Cohen. 2006. Biomedical Language Processing: Perspective Whats Beyond PubMed? *Mol Cell.*, 21(5):589–594.

C.-Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out.*, pages 74–81, Barcelona, Spain.

I. Mani. 2001. *Automatic summarization*. Jonh Benjamins Publishing Company.

R. Mihalcea and P. Tarau. 2004. TextRank - Bringing order into text. In *Proceedings of the Conference EMNLP 2004*, pages 404–411.

R. Navigli and P. Velardi. 2005. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1075–1086.

S. Nelson, T. Powell, and B. Humphreys. 2002. The Unified Medical Language System (UMLS) Project. In Allen Kent and Carolyn M. Hall, editors, *Encyclopedia of Library and Information Science*. Marcel Dekker, Inc.

L. Plaza, A. Díaz, and P. Gervás. 2008. Concept-graph based biomedical automatic summarization using ontologies. In *TextGraphs '08: Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, pages 53–56.

L.H. Reeve, H. Han, and A.D. Brooks. 2007. The use of domain-specific concepts in biomedical text summarization. *Information Processing and Management*, 43:1765–1776.

M. Weeber, J. Mork, and A. Aronson. 2001. Developing a Test Collection for Biomedical Word Sense Disambiguation. In *Proceedings of AMIA Symposium*, pages 746–50, Washington, DC.

I. Yoo, X. Hu, and I-Y. Song. 2007. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. *BMC Bioinformatics*, 8(9).