Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Resolving ambiguity in biomedical text to improve summarization

Laura Plaza [a,*], Mark Stevenson [b], Alberto Díaz [a]

[a] Dpto. de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, C/ Profesor José García Santesmases s/n, 28040 Madrid, Spain
[b] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom

ARTICLE INFO

Article history:
Received 10 August 2010
Received in revised form 24 June 2011
Accepted 15 September 2011
Available online xxxx

Keywords:
Biomedical summarization
Word sense disambiguation
WSD
Unified medical language system
UMLS
MetaMap

ABSTRACT

Access to the vast body of research literature that is now available on biomedicine and related fields can be improved with automatic summarization. This paper describes a summarization system for the biomedical domain that represents documents as graphs formed from concepts and relations in the UMLS Metathesaurus. This system has to deal with the ambiguities that occur in biomedical documents. We describe a variety of strategies that make use of MetaMap and Word Sense Disambiguation (WSD) to accurately map biomedical documents onto UMLS Metathesaurus concepts. Evaluation is carried out using a collection of 150 biomedical scientific articles from the BioMed Central corpus. We find that using WSD improves the quality of the summaries generated.

1. Introduction and background

A vast amount of literature on biomedicine and related fields is now available and growing at an increasing rate (Hunter & Cohen, 2006). Access to the information it contains is necessary for researchers and has also been shown to be useful for both health professionals and consumers (Lau & Coiera, 2008; Westbrook, Coiera, & Gosling, 2005). However, the amount of information available is now so large that tools are required in order to access it practically (Cohen & Hersh, 2005; Zweigenbaum, Demner-Fushman, Yu, & Cohen, 2007). Text summarization systems can improve this access (Hunter & Cohen, 2006; Reeve, Han, & Brooks, 2007). When no author's abstract is available, researchers can use summaries to determine whether a scientific article is of interest without having to read the entire document (Mani, 1999, 2001; Moens, 2000). Automatic summarization systems may be also used to assist scientists to write abstracts. Physicians can use summaries to identify treatment options, reducing the diagnosis time (Brooks & Sulimanoff, 2002). Reeve et al. (2007) states that there are two reasons for generating summaries from a full-text source, even when the author has created an abstract: (1) the abstract may not include relevant content from the full-text, and (2) there is no single "ideal" summary that meets the information needs of all users. Moreover, automatic summaries have been shown to improve indexing and categorization of biomedical literature, when used instead of the articles' abstracts (Gay, Kayaalp, & Aronson, 2005).

Summarization systems usually work with *text-level representations* of the document which consist of information that can be directly extracted from the document itself (Erkan & Radev, 2004; Mihalcea & Tarau, 2004). Studies have also demonstrated the benefit of richer *conceptual representations* (Fiszman, Rindflesch, & Kilicoglu, 2004; Plaza, Díaz, & Gervás, 2008), which represent documents using concepts instead of words. The representations may be enriched with semantic relations between the concepts (i.e. synonymy, hypernymy, homonymy or co-occurrence) to improve the quality of the summaries.

* Corresponding author. Tel.: +34 625638290; fax: +34 913947529.
E-mail addresses: lplazam@fdi.ucm.es (L. Plaza), m.stevenson@dcs.shef.ac.uk (M. Stevenson), albertodiaz@fdi.ucm.es (A. Díaz).

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.09.005

The Unified Medical Language System (UMLS) (Nelson, Powell, & Humphreys, 2002) has proved to be a useful knowledge source for summarization in the biomedical domain (Fiszman et al., 2004; Plaza et al., 2008; Reeve et al., 2007). When the UMLS is used, the vocabulary of the document being summarized has to be mapped onto the concepts it contains. This is made difficult by lexical ambiguity, the fact that words can have multiple meanings depending on the context in which they appear. Although it is often believed that technical domains contain less ambiguity than general ones (Farghlay & Hedin, 2003; Gale, Church, & Yarowsky, 1992), biomedical text has been shown to be highly ambiguous (Weeber, Mork, & Aronson, 2001). For example, the term “cold” is associated with several possible meanings in the UMLS Metathesaurus including ‘common cold’, ‘cold sensation’, ‘cold temperature’ and ‘cold therapy’.

The majority of biomedical summarizers that employ the UMLS Metathesaurus use MetaMap (Aronson, 2001) to translate the text into UMLS concepts (Fiszman et al., 2004; Reeve et al., 2007) but do not attempt to resolve ambiguities when MetaMap returns multiple concepts. However, selecting the wrong meaning for ambiguous terms may affect the quality of the summaries generated.

This paper describes the application of various strategies for selecting UMLS concepts from the MetaMap output to improve a state-of-art biomedical summarization system. The summarizer (Plaza et al., 2008) is a graph-based method that uses the UMLS Metathesaurus to create conceptual representations. Strategies for selecting concepts from MetaMap include using Word Sense Disambiguation (WSD) (Agirre & Edmonds, 2006) to attempt to determine the meaning of words by examining their context. We find that using WSD improves the quality of the summaries generated.

The next section describes related work on summarization and WSD and also introduces the resources employed by the summarization and WSD systems used in this work. Section 3 describes our concept-based summarization algorithm. Section 4 presents the different WSD algorithms and strategies that have been tested to assign concepts from the UMLS. Section 5 describes the experimental environment of the study. Section 6 reports the results of the experiments and discusses these results. The final section provides concluding remarks and suggests future lines of work.

2. Related work

2.1. UMLS and MetaMap

The Unified Medical Language System (UMLS) (Nelson et al., 2002) is a collection of controlled vocabularies related to biomedicine and contains a wide range of information that can be used for Natural Language Processing (NLP). The UMLS comprises of three parts: the Specialist Lexicon, the Metathesaurus and the Semantic Network.

The **Specialist Lexicon** is a database of lexicographic information for use in NLP tasks that consists of a set of lexical entries with one entry for each spelling or set of spelling variants in a particular part of speech.

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’.

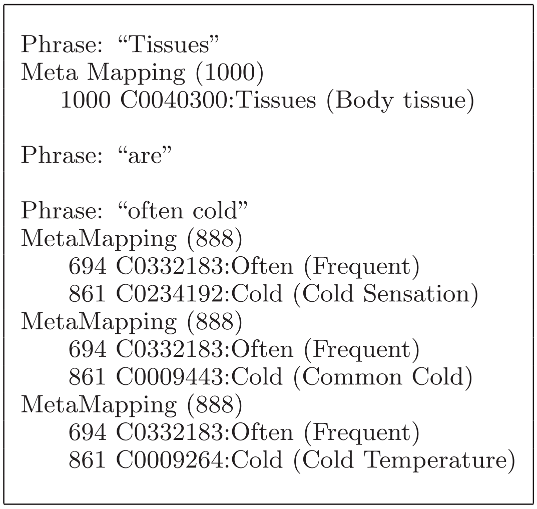
The Metathesaurus comprises of several tables containing information about CUIs. These include the `MRREL` and `MRHIER` tables. The `MRREL` table lists relations between CUIs found in the various sources that are used to form the Metathesaurus. This table lists a range of different types of relations, including `child`, `parent`, `can be qualified by`, `related and possibly synonymous` and `other related`. For example, the `MRREL` table states that the concepts C0009443 ‘Common Cold’ and C0027442 ‘Nasopharynx’ are connected via the `other related` relation.

The `MRHIER` table in the Metathesaurus lists the hierarchies in which each CUI appears, and lists the entire path to the root of each hierarchy for the CUI.

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 concept 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.

The **MetaMap** program (Aronson, 2001) maps biomedical text to concepts in the Metathesaurus. The semantic type for each concept mapping is also returned. MetaMap employs a knowledge intensive approach that uses the Specialist Lexicon in combination with lexical and syntactic analysis to identify noun phrases in text. Matches between noun phrases and Metathesaurus concepts are computed by generating lexical variations and allowing partial matches between the phrase and concept. The possible UMLS concepts are assigned scores based on the closeness of the match between the input noun phrase and the target concept. Fig. 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 multiple CUIs with equal scores for *cold* (C0234192, C0009443 and C0009264). Weeber et al. (2001) estimated that around 11% of the phrases in Medline abstracts are mapped onto multiple CUIs.



```
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)
```

Fig. 1. MetaMap output for the phrase *Tissues are often cold*.

2.2. Summarization of biomedical text

Summarization has been an active area within NLP research since the 1950s. A variety of approaches have been proposed (see Mani (2001) for a review). We focus on graph-based methods here.

Graph-based approaches represent the document as a graph. When text-level representations are used the nodes represent text units (i.e. words, sentences or paragraphs) and the edges represent cohesion relations or similarity measures between these units. The best-known work in the area is LexRank (Erkan & Radev, 2004). It assumes a fully connected and undirected graph in which 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) describe a related method in which the similarity among sentences is measured in terms of word overlaps.

However, text-level representations do not exploit the semantic relations among the words in the text (i.e. synonymy, homonymy or co-occurrence). For example, they are unable to make use of the fact 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 solved using conceptual representations of the document and domain-specific resources that contain information about the semantic relations between concepts.

For example, Reeve et al. (2007) use the UMLS Metathesaurus and adapt the lexical chaining approach (Barzilay & Elhadad, 1997) to deal with concepts instead of terms. Yoo, Hu, and Song (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. BioSquash (Shi et al., 2007) is a question-oriented extractive system for biomedical multi-document summarization. It constructs a semantic graph that contains concepts of three types: ontological concepts (general ones from WordNet and specific ones from the UMLS), named entities and noun phrases. It applies a WSD algorithm to select the correct senses in WordNet although they do not describe how they dealt with ambiguity in the UMLS.

2.3. WSD of biomedical text

The aim of WSD is to resolve lexical ambiguities by identifying the correct meaning of a word based on its context. WSD is regarded as an important stage in text processing (Agirre & Edmonds, 2006; Ide & Véronis, 1998; Navigli, 2009). The majority of approaches have explored the problem in a domain-independent setting, although several researchers have developed systems specifically intended to resolve the ambiguities that are found in the biomedical domain (see Schuemie, Kors, & Mons (2005) for a review).

The most popular approaches to WSD in the biomedical domain are based on supervised learning, for example Joshi, Pedersen, and Maclin (2005), Liu, Teller, and Friedman (2004), McInnes, Pedersen, and Carlis (2007), Stevenson, Guo, Gaizauskas, and Martinez (2008) and Savova et al. (2008). Previous evaluations of domain-independent WSD (Mihalcea & Tarau, 2004; Pradhan, Loper, Dligach, & Palmer, 2007) have suggested that supervised approaches perform better than alternative approaches. However, they require labeled examples which are often unavailable and can be impractical to create.

Humphrey, Rogers, Kilicoglu, Demner-Fushman, and Rindflesch (2006) describe a supervised approach to WSD, JDI, in the biomedical domain that makes use of *Journal Descriptors* to create labeled examples. Journal Descriptors are a mapping from the journals indexed in Medline to a set of 127 broad topics such as *Reproduction*, *Biomedical Engineering* and *Cardiology*. Medline is used to create a model for each Journal Descriptor by mapping each abstract onto its manually added descriptors and identifying the words that characterize it. Semantic types are then modeled by using UMLS to identify terms associated with them and mapping these onto Journal Descriptors. Disambiguation is carried out by examining the context of the ambiguous term to identify the semantic type it matches more closely. One shortcoming of this algorithm is that it is not able to disambiguate between meanings with the same semantic type. The JDI algorithm is included in recent versions of MetaMap and can be used to select between mappings when there is more than one possible CUI, as shown in Fig. 1. Humphrey et al. (2006) report precision of 0.7860 for JDI when evaluated against a set of 45 ambiguous terms from the NLM-WSD corpus (Weeber et al., 2001).

Knowledge-based approaches to WSD are an alternative to supervised learning that do not require manually-tagged data and have recently been shown to compete with supervised systems in terms of performance (Ponzetto & Navigli, 2010). Graph-based methods are now widely used for knowledge-based WSD (Agirre & Soroa, 2009; Navigli & Lapata, 2007; Sinha & Mihalcea, 2007; Tsatsaronis, Vazirgiannis, & Androutsopoulos, 2007). These methods represent the knowledge base as a graph which is then analyzed to identify the meanings of ambiguous words. An advantage of this approach is that the entire knowledge base can be used during the disambiguation process by propagating information through the graph.

One such method is Personalized PageRank (Agirre & Soroa, 2009) which makes use of the PageRank algorithm used by internet search engines (Brin & 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) use WordNet as the lexical knowledge base and create graphs using the entire WordNet hierarchy. The ambiguous words in the document are added as nodes to this graph and directed links 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`. 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. 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 domain-independent WSD and report an accuracy of around 58% on standard evaluation data sets. Agirre, Sora, and Stevenson (2010) applied Personalised PageRank to the biomedical domain using the UMLS instead of WordNet and report precision of 0.681 on a set of 49 terms from the NLM-WSD corpus.

3. Summarizer

This section describes the graph-based conceptual summarization system used in this paper. The summarizer identifies the *n* most relevant sentences in a document using a 5 steps process: (1) preprocessing, (2) concept identification, (3) document representation, (4) concept clustering and topic recognition, and (5) sentence selection. Each step is now discussed in detail.

3.1. Preprocessing

Before starting with the summarization process itself, a preliminary step is undertaken in order to prepare the document for the subsequent steps. This preprocessing involves the following actions:

1. Portions of the document that are not considered suitable for including in the summary are removed: *Competing interests*, *Acknowledgments* and *References* sections, tables, figures and section headings.
2. If the document includes an *Abbreviations* section the abbreviations and their corresponding expansions are extracted from it. Any occurrences of the abbreviation in the document are then replaced with the expansion. For example, if the *Abbreviations* section defines "angiotensin converting enzyme" as the expansion of ACE for a particular document and that document contains the phrase "less than 50% received ACE inhibitor" then that phrase would become "less than 50% received angiotensin converting enzyme inhibitor".
3. Expansions for abbreviations not defined in the document's abbreviation section are identified automatically from the document using a publicly available implementation[1] of the approach described by Schwartz and Hearst (2003). Abbreviations are then substituted with their expansions in the same way.
4. The text in the body section is split into sentences using GATE[2] (*General Architecture for Text Engineering*).
5. Stop words[3] are removed since they are not useful in discriminating between relevant and irrelevant sentences.

[1] http://biotext.berkley.edu/software.html.

[2] http://www.gate.ac.uk.

[3] A list of stop words from Medline are used. http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#Stopwords.

Please cite this article in press as: Plaza, L., et al. Resolving ambiguity in biomedical text to improve summarization. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.09.005

3.2. Concept identification

The next stage is to map the terms in the document to concepts from the UMLS Metathesaurus and semantic types from the UMLS Semantic Network. The MetaMap program is run over the text in the body section of the document. Section 4 describes different strategies to select concepts when MetaMap is unable to return a single CUI for a word.

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*.

3.3. Document representation

The next step is to construct a graph representing the document. We begin by creating a *sentence graph* for each sentence in the document. This is created using the UMLS concepts identified within the sentence, extracting the complete hierarchy of hypernyms for each concept and merging the hierarchies to construct a single graph. The two upper levels of these hierarchies are removed since they represent concepts with excessively broad meanings and may introduce noise to later processing.

The sentence graphs are then merged to create 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 with different sets of relations and found that the best performance was obtained using 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, `other related` relations from the `MRREL` table and `associated with` relations from the `SRSTR` table (see Section 2.1). Only `associated with` and `other related` relations that link leaf vertices are added to the document graph.

Fig. 2 shows an example graph for a simplified document consisting of the sentences `s1` and `s2`:

`s1` *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.*

`s2` *The trial was carried out in two groups: the first group taking doxazosin, and the second group taking chlorthalidone.*

Finally, each edge is assigned a weight in [0, 1] as shown in Eq. (1). The weight of an edge *e* representing an `is_a` relation between two vertices, v_i and v_j (where v_i is a parent of v_j), is calculated as the ratio of the depth of v_i to the depth of v_j from the root of their hierarchy. The weight of an edge representing any other relation (i.e. `associated with` and `other related`) between pairs of leaf vertices is always 1.

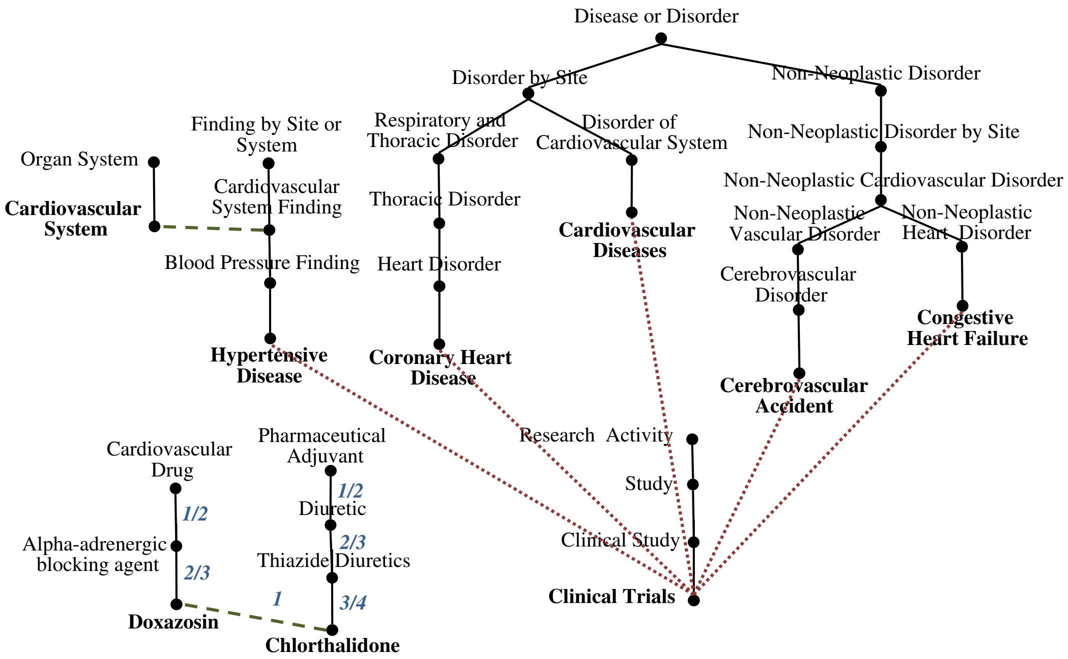


Fig. 2. Example of a simplified document graph from sentences `s1` and `s2`. Continuous lines represent *hypernymy* relations, dashed lines represent *other related* relations and dotted lines represent *associated with* relations. The edges of a portion of this graph have been labeled with their weights.

Please cite this article in press as: Plaza, L., et al. Resolving ambiguity in biomedical text to improve summarization. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.09.005

$$weight(e, v_i, v_j) = \beta \tag{1}$$

$$\text{where}\begin{cases} \beta = \frac{depth(v_i)}{depth(v_j)} & \text{if } e \text{ represents an } \texttt{is\_a} \text{ relation} \\ \beta = 1 & \text{otherwise} \end{cases}$$

This is shown in Fig. 2, where the is_a link between concepts 'Diuretic' and 'Thiazide Diuretics' is assigned the weight $\frac{2}{3}$ since 'Diuretic' is ranked second in its hierarchy and 'Thiazide Diuretics' is ranked third in this same hierarchy. In this same graph, the other related link between concepts 'Doxazosin' and 'Chlorthalidone' is assigned the weight 1.

3.4. Concept clustering and topic recognition

A *degree-based clustering* method (Erkan & Radev, 2004) is applied to the document graph. This identifies sets of concepts that are strongly related in meaning and it is assumed that each set represents a different topic in the document.

We hypothesize that the document graph is an instance of a *scale-free network* (Barabasi & Albert, 1999). These are complex networks in which some nodes are highly connected to other nodes in the network while the remaining nodes are relatively unconnected. The highly connected nodes are often called *hubs*.

The *salience* of each vertex in the graph (Yoo et al., 2007) is then computed. The salience of a vertex, v_i, is defined as the sum of the weights of the edges that are connected to it. This is shown in Eq. (2) where $connect(e, v_i, v_j)$ denotes that the edge e connects vertices v_i and v_j. Salience ranks the nodes according to their structural importance in the graph.

$$salience(v_i) = \sum_{e|\exists v_j \wedge connect(e, v_i, v_j)} weight(e) \tag{2}$$

The n vertices with a highest salience are labeled as *Hub Vertices*. This value has been empirically set to 5% of the number of vertices in the document graph. A clustering algorithm groups the hub vertices into *Hub Vertices Sets (HVS)*. These can be interpreted as set of concepts strongly related in meaning and represent the centroids of the clusters. To construct the HVS, the clustering algorithm first identifies the pairs of hub vertices that are most closely connected and merges them into a HVS. Then, for each pair of HVS, the algorithm checks if the internal connectivity of the vertices they contain is lower than the connectivity between them. If it is the two 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.

3.5. Sentence selection

The final step of the summarization process consists of selecting the sentences for the summary. A non-democratic voting mechanism (Yoo et al., 2007) is used to compute the similarity between each sentence graph and cluster. Each vertex, v_k, within a sentence graph, S_j, assigns a vote $w_{k,j,i}$ to a cluster C_i if the vertex belongs to that cluster's HVS, half a vote if the vertex belongs to the cluster 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 shown in Eq. (3).

$$similarity(C_i, S_j) = \sum_{v_k | v_k \in S_j} w_{k,j,i} \tag{3}$$

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

We select the sentences for the summary based on their similarities to the clusters. We compute a single score for each sentence as the sum of its similarity to each cluster normalized by the cluster's size (Eq. (4)). The N sentences with highest scores are then selected for the summary, where N depends on the summary compression rate.

$$Score(S_j) = \sum_{C_i} \frac{similarity(C_i, S_j)}{|C_i|} \tag{4}$$

This clustering method usually produces a single large cluster together with a variable number of small clusters. The large cluster contains the concepts related to the central topic of the document, while the others include concepts related to secondary information. The scoring function selects most of the sentences from the most populated cluster and also includes some sentences from other clusters when the sentences assign high scores to them. Therefore, in addition to the information related to the central topic, this function allows us to consider other secondary or "satellite" information that might be relevant to the user.

Alternative heuristics for sentence selection were explored in previous work (Plaza et al., 2008).

4. WSD strategies for concept identification

Since our summarization system is based on the UMLS it is important to be able to accurately map the documents onto Metathesaurus concepts. The example in Section 2.1 shows that MetaMap does not always select a single concept and it is therefore necessary to have some method for choosing between the ones that are returned. We compare several alternative approaches for concept selection, including ones that make use of WSD.

4.1. First mapping

The first approach used is to take the first mapping returned by MetaMap when it returns more than one concept. No attempt is made to resolve ambiguities. This approach is adopted in previous works (Plaza et al., 2008; Reeve et al., 2007). However, since the order of equally scored concepts returned by MetaMap is not informative, this strategy is essentially the same as choosing a concept at random from the set of those returned.

4.2. WSD using PPR

A version of the Personalized PageRank algorithm[4] (see Section 2.3) is used to disambiguate the output of MetaMap and the concept chosen for each term used to create the graph.

The Personalized PageRank algorithm was adapted to assign concepts from the UMLS Metathesaurus following the process described by Agirre et al. (2010). The UMLS is converted into a graph in which the concepts are the nodes and the edges are derived from the `MRREL` table. All relations in this table are included in the graph. The output from MetaMap is used to provide the list of possible concepts 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 for disambiguation. The concepts selected by the disambiguation algorithm are then used to create the document graph.

4.3. WSD using JDI

The JDI algorithm (see Section 2.3) was used as an alternative WSD approach. The implementation included with MetaMap was used.[5] The concepts returned by MetaMap when the JDI algorithm is applied are used to create the document graph. When the possible concepts of a term share the same semantic type the JDI algorithm may fail to return a single sense. When this happens the first mapping returned by MetaMap is selected.

4.4. All mappings

Instead of using WSD, all candidate CUIs are used to build the document graph. No attempt is made to resolve ambiguity when there are multiple possible CUIs.

4.5. WSD using weighted mappings

The "All Mappings" strategy made use of all concepts returned by MetaMap and considered them all to be equally important. The final strategy also uses all concepts returned by MetaMap but weight them using the output of a WSD algorithm. The aim of this strategy is to reduce the effect of WSD errors when the document graph is created.

The function for computing the salience of the vertices in the document graph (Eq. (2)) is modified to assign greater weight to the concept selected by the WSD algorithm. The modified function is shown in Eq. (5), where *M* is the number of possible CUIs returned by MetaMap for a given ambiguous term and *salience*(v_i) was defined in Eq. (2).

$$weighted\_salience(v_i) = salience(v_i) \times \alpha_i \tag{5}$$

$$\text{where} \begin{cases} \alpha_i = 2 \times \frac{1}{M+1} & \text{if } v_i \in WSD\_output \\ \alpha_i = \frac{1}{M+1} & \text{if } v_i \notin WSD\_output \end{cases}$$

5. Experimental method

5.1. Evaluation collection

The most common approach to evaluating automatically generated summaries of a document (also known as *peers*) is to compare them against manually-created summaries (*reference* summaries) and measure the similarity between their con-

[4] We use a publicly available implementation of the Personalized Page Rank algorithm (http://ixa2.si.ehu.es/ukb/).

[5] The JDI algorithm is invoked in MetaMap using the `-y` flag.

Table 1
ROUGE scores for the different configurations of the summarizer. The best score for each metric is indicated in bold font. Systems are sorted by decreasing R-2 score.

Summarizer	R-1	R-2	R–W	R-SU4
WSD weighted mappings	**0.7908**	**0.3590**	**0.2026**	**0.3362**
All mappings	0.7873	0.3577	0.2014	0.3303
WSD JDI	0.7874	0.3560	0.2017	0.3300
WSD PPR-W2W	0.7826	0.3542	0.1999	0.3295
WSD PPR	0.7737	0.3419	0.1937	0.3178
First mapping	0.7516	0.3300	0.1924	0.3122
Lead baseline	0.6483	0.2566	0.1621	0.2646

tent. The more content that is shared between the peer and reference summaries, the better the peer summary is assumed to be. One of the main drawbacks of this type of evaluation is the difficulty in obtaining reference summaries, which have to be written by humans. To the authors' knowledge no corpus of reference summaries exists for biomedical documents. However, most scientific papers include an abstract (i.e. the author's summary) which can be used as a reference summary for evaluation.

Our approach was evaluated using a collection of 150 documents randomly selected from the BioMed Central corpus for text mining research.[6] This collection is large enough to allow significant evaluation results (Lin, 2004a). The abstracts for the papers were used as reference summaries. Abstracts in the BioMed Central corpus are frequently structured and contain sections such as background, method, results and conclusion. However, we do not make use of this structure in our approach.

5.2. Evaluation metrics

ROUGE (Lin, 2004b) is used to evaluate the summaries by comparing them with the human abstracts for each article in the evaluation corpus. ROUGE is a commonly used evaluation method for summarization which uses the proportion of n-grams between a peer and one or more reference summaries to estimate the content that is shared between them. The ROUGE metrics produce a value in $[0,1]$, where higher values are preferred, since they indicate a greater content overlap between the peer and model summaries. The following ROUGE metrics were used: ROUGE-1 (**R-1**), ROUGE-2 (**R-2**), ROUGE-SU4 (**R-SU4**) and ROUGE-W (**R-W**). R-1 and R-2 compute the number of unigrams and bigrams, respectively, that are shared by the peer and reference summaries. R-SU4 measures the overlap of skip-bigrams between the peer and reference summaries, allowing a skip distance of 4. Finally, R–W computes the union of the longest common subsequences between the peer and the reference summaries by taking the presence of consecutive matches into account.

ROUGE metrics assess the content of the summaries but do not account for text readability. At present readability of automatic summaries is still evaluated manually by experts. For instance, the assessors of the Text Analysis Conferences (TAC)[7] use a list of linguistic quality criteria (i.e. grammaticality, non-redundancy, referential clarity, focus, and structure and coherence), and manually assign a score to each summary depending on the extent to which they meet each criteria. However, manual evaluation of this type is expensive, time consuming and difficult to replicate. As a consequence the research community has explored strategies for automatically evaluating the readability of a summary. For instance, Lapata and Barzilay (2005) introduce a linguistically rich model of local coherence to automatically evaluate text coherence for machine-generated texts that incorporates both syntactic and semantic aspects. Pitler, Louis, and Nenkova (2010) analyze how well different types of automatically computed features (such as word choice, reference form and local coherence) can rank automatic summaries according to the five linguistic quality criteria used in the TAC conferences. Other attempts to automatically measure the quality of a summary can be found in (Barzilay & Lapata, 2005; Pitler & Nenkova, 2008; Vadlapudi & Katragadda, 2010). However, research in automatic evaluation of readability is still very preliminary and no standard approach has been adopted by the research community.

A second drawback of ROUGE metrics is that they use lexical matching instead of semantic matching. Therefore, peer summaries that are worded differently but carry the same semantic information may be assigned different ROUGE scores. This phenomenon is known as paraphrase. Zhou, Lin, Munteanu, and Hovy (2006) present ParaEval, a preliminary approach to automatically evaluate summaries using paraphrases.

In contrast, the main advantages of ROUGE are its simplicity, replicability and, particularly, the high correlation between the scores it produces and those provided by human judges from previous Document Understanding Conferences (Lin, 2004b).

5.3. Experiments

Automatic summaries are generated by selecting sentences until the summary is 30% of the original document size. This choice of summary size was based on the heuristic that a summary should be between 15% and 35% of the size of the source

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

[7] http://www.nist.gov/tac/.

text (Hovy, 2005). The relatively large size within this range was chosen because the documents used for the experiments (i.e. scientific articles) are rich in information.

Six different types of summaries are created using various strategies for concept identification (Section 4): (1) first mapping returned by MetaMap (**First Mapping**), (2) WSD using standard Personalized PageRank (**WSD PPR**), (3) WSD using word-to-word variant of Personalized PageRank (**WSD PPR-W2W**), (4) WSD using JDI (**WSD JDI**), (5) all mappings generated by MetaMap (**All Mappings**) and (6) all mappings generated by MetaMap weighted with WSD output (**WSD Weighted Mappings**). In addition, a baseline summarizer (**Lead Baseline**) was also implemented. This baseline generates summaries by selecting the first *N* sentences from each document until the document reaches a length of 30% of the original document size.

A Wilcoxon Signed Ranks Test with a 95% confidence interval is used to test statistical significance of the results.

6. Results and analysis

ROUGE scores for the summaries generated using the different strategies for concept identification are shown in Table 1. Results show that the JDI algorithm outperforms the Personalized PageRank approaches and consequently JDI is used for the "WSD Weighted Mappings" approach.

Using WSD improves the average ROUGE scores for the summarizer compared to the "First Mapping" baseline. This improvement is observed for all approaches to WSD. The "standard" (i.e. WSD PPR) version of the Personalized PageRank disambiguation algorithm significantly improves R-1 and R-2 metrics while the "word to word" variant (i.e. WSD PPR-W2W) significantly improves all ROUGE metrics. Performance using "word to word" PPR is also higher than standard PPR for all ROUGE metrics, and these differences are significant for all ROUGE metrics. Results using the JDI algorithm (i.e. WSD JDI) are better than those achieved by either Personalized PageRank variant. However, the improvement with respect to WSD PPR-W2W is not statistically significant.

Results using the simple "All Mappings" approach are comparable to those obtained using the best WSD algorithm, i.e. JDI. This approach slightly improves upon the results of the JDI algorithm for R-2 and R-SU4 metrics, but does not perform as well for R-1 and R-W.

The best results are obtained using the "WSD Weighted Mappings" approach. This strategy achieves the best result for all ROUGE metrics, increasing the R-1 metric by 3.9% compared to the first mapping baseline. This performance is significantly better than when the first mapping and PPR approaches are used for all ROUGE metrics. Although, the improvement over PPR-W2W, JDI and "All Mappings" is not significant.

Finally, it must be noted that performance of all variants of our graph-based summarizer is considerably better than the lead baseline.

6.1. Analysis

The results presented above demonstrate that using WSD improves the performance of our summarizer compared to the first mapping baseline. The WSD algorithms identify the concepts that are being referred to in the documents more accurately which leads to the creation of a graph that better reflects the content of the document. As a result, the clustering method is better able to identify the topics covered in the document and the information in the sentences selected for the summary is closer to the model summaries. However, this improvement is less than expected and this is probably due to errors made by the WSD system (see Section 2.3).

On the other hand, the differences between the summaries generated using the 'word to word' PPR and the JDI algorithms are not significant; the second achieving an improvement of less than 0.5% for ROUGE-1. Both algorithms use a different semantic knowledge source (i.e. PPR uses the concepts and relations from the UMLS Metathesaurus, while JDI uses the information about the semantic types in the UMLS Semantic Network to which the different concepts belong to), but both yield to a similar improvement in the summarization results. This seems to indicate that both the Metathesaurus and the Semantic Network provide useful information for biomedical WSD, and combining the information from both sources will probably reduce the disambiguation errors.

Performance improves when all possible concepts for ambiguous terms are used to build the document graph and no WSD is used. This suggests that the performance of our summarizer is not as sensitive to ambiguity as it is to erroneous disambiguation. If WSD is incorrect then the correct concept is never passed to the summarization algorithm, making accurate representation of the document difficult. The summarizer itself may also be performing WSD implicitly. The graph-based algorithm for summarization algorithm (Section 3) is similar to Personalized PageRank, a graph-based WSD algorithm (Section 4.2). When all possible concepts for each term are included in the graph then the correct ones would be strongly connected to other concepts in the graph while the incorrect ones will be strongly disconnected. The topology of the graph will exhibit a more marked scale-free distribution, since a relatively small set of nodes will present a great number of connections with other nodes while the remaining nodes will present few connections. Consequently the correct concepts will be given a higher salience than the others in the concept clustering step and their influence in the sentence selection step will be greater.

Finally, the best results are achieved when all candidate concepts are used to build the document graph with extra importance attached to those identified using WSD. This shows that the information provided by the WSD algorithm is useful for

Please cite this article in press as: Plaza, L., et al. Resolving ambiguity in biomedical text to improve summarization. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.09.005

Table 2
Examples of summaries generated using different WSD strategies (WSD weighted mappings and first mapping), the lead baseline and the authors' abstract.

WSD weighted mappings: [1] Elbow dislocation is a common injury, postero-lateral dislocation being the commonest pattern of injury. [2] We present an unusual case of an open antero-lateral dislocation of the elbow, which was not associated with any vascular or neural injury. [3] A 34 year female dance instructor presented to the Accident and Emergency department having fallen onto her right elbow sustaining an open dislocation of her elbow. [4] There was almost complete disruption of all the muscular and ligamentous attachments to the distal humerus and the proximal radius and ulna. [5] The joint was relocated, the wound cleaned with saline lavage and primarily closed without formal repair of muscular and ligamentous structures. [6] *The functional recovery was complete in about six months, the patient regaining full range of elbow movement.* [7] Posterior elbow dislocation is due to a combined valgus and external rotatory stress to the semiflexed elbow, resulting in a bilateral ligamentous injury. [8] Anterior elbow dislocations occur most often as a fracture-dislocation in which the distal humerus is driven through the olecranon. [9] Most authors recommend accelerated functional treatment for simple elbow dislocations, as long periods of immobilisation have not been found to be of any benefit.

First mapping: [1] Elbow dislocation is a common injury, postero-lateral dislocation being the commonest pattern of injury. [2] We present an unusual case of an open antero-lateral dislocation of the elbow, which was not associated with any vascular or neural injury. [3] A 34 year female dance instructor presented to the Accident and Emergency department having fallen onto her right elbow sustaining an open dislocation of her elbow. [4] *The uninjured brachial artery, and ulnar and median nerves were all visualized.* [5] There was almost complete disruption of all the muscular and ligamentous attachments to the distal humerus and the proximal radius and ulna. [6] The joint was relocated, the wound cleaned with saline lavage and primarily closed without formal repair of muscular and ligamentous structures. [7] Posterior elbow dislocation is due to a combined valgus and external rotatory stress to the semiflexed elbow, resulting in a bilateral ligamentous injury. [8] Anterior elbow dislocations occur most often as a fracture-dislocation in which the distal humerus is driven through the olecranon. [9] Most authors recommend accelerated functional treatment for simple elbow dislocations, as long periods of immobilisation have not been found to be of any benefit.

Lead baseline: [1] Elbow dislocation is a common injury, postero-lateral dislocation being the commonest pattern of injury. [2] Open dislocations are infrequent, often associated with damage to the neurovascular structures. [3] These injuries and various ways of dealing with them have been widely described in the literature. [4] We present an unusual case of an open antero-lateral dislocation of the elbow, which was not associated with any vascular or neural injury. [5] A 34 year female dance instructor presented to the Accident and Emergency department having fallen onto her right elbow sustaining an open dislocation of her elbow. [6] On admission there was a palpable radial pulse and full sensation in her forearm and hand. [7] Radiographs revealed this to be an anterolateral elbow dislocation. [7] Surgical exploration was undertaken. [8] The uninjured brachial artery, and ulnar and median nerves were all visualized. [9] There was almost complete disruption of all the muscular and ligamentous attachments to the distal humerus and the proximal radius and ulna.

Abstract: [1] Open dislocations are infrequent, often associated with damage to the neuro vascular structures. [2] We present an unusual case of an open antero-lateral dislocation of the elbow, which was not associated with any vascular or neural injury. [3] A 34 year female dance instructor sustained an open dislocation of her elbow. [4] Surgical exploration was undertaken. [5] No major neurovascular injury was present. [6] There was almost complete disruption of all the muscular and ligamentous attachments to the distal humerus and the proximal radius and ulna, which were not formally repaired during surgery. [7] The elbow was found to be very unstable, and was placed in a back slab. [8] The functional recovery was complete in about six months, the patient regaining full range of elbow movement. [9] Elbow dislocations without associate fractures are adequately treated by manipulation and reduction, in spite of the almost complete disruption of the soft tissues around the joint.

summarization. However, the WSD is not accurate enough to be relied upon alone and it is important that it is applied in a suitable way, that is weighting likely concepts rather than removing some from the document graph.

The use of WSD to improve summarization seems similar to previous work which showed that WSD could improve Information Retrieval performance but only when the disambiguation was accurate enough (Sanderson, 1994).

To end with, Table 2 shows the automatic summaries generated by the best disambiguation strategy (“WSD Weighted Mappings”) and the worst strategy (“First Mapping”), respectively, for a document from the BioMed Central corpus on an open dislocation of the elbow. The lead baseline and the abstract of the paper are also shown. The “WSD Weighted Mappings” and “First Mapping” summaries differ in only 1 of the 9 sentences selected (sentences [6] and [4], respectively). The reason for this small difference seems to be that both strategies agree when assigning meanings to the concepts that represent the topics of the document, particularly to those belonging to the HVS or centroids of the clusters. This agreement is achieved despite the fact there is nearly 10% disagreement in the concept mappings for both strategies.

7. Conclusion

This paper explores the integration of WSD algorithms with a graph-based approach to biomedical summarization. The summarizer represents the document as a graph, where the nodes are concepts from the UMLS Metathesaurus and the links are relations between them. This produces a richer representation than the one provided by traditional term-based models. Our approach relies on accurate mapping of the document being summarized onto the concepts in the UMLS Metathesaurus. We found that the commonly used approach of choosing the first mapping from MetaMap could be improved by choosing the concept identified by a WSD system. However, we also found the simple approach of choosing all concepts returned by MetaMap produced performance comparable to the best WSD approach. The best performance was obtained when all possible concepts from MetaMap were used to build the document graph and information from a WSD system used to weight them.

Overall, the results presented in this paper suggest that WSD benefits summarization performance and that this benefit is limited by the accuracy of the disambiguation. Reducing the errors made by WSD systems may produce better summaries and this will be explored.

Our future aim is to improve the summarizer to the extent that the abstract writing process can be completely automated. Improvements will involve analyzing the structure of biomedical scientific papers and their abstracts in order to weight the sentences for extraction according to the section in which they appear; and adapting the method to produce query-driven summaries, so that the summaries are biased toward the user's information needs. We also aim to extend the method to produce multi-document summaries.

We are also interested in exploring and using new automatic evaluation strategies. In particular, we would like to measure readability quality (Pitler et al., 2010) as well as informativeness without making use of human model summaries (Louis & Nenkova, 2008; Saggion, Torres-Moreno, Cunha, & SanJuan, 2010).

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 is grateful for the support provided by the Engineering and Physical Sciences Research Council (Grant EP/D069548/1).

References

Agirre, E., & Edmonds, P. (Eds.). (2006). *Word sense disambiguation: Algorithms and applications*. Springer.
Agirre, E., Sora, A., & Stevenson, M. (2010). Graph-based word sense disambiguation of biomedical documents. *Bioinformatics, 26*, 2889–2896.
Agirre, E., & Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009)* (pp. 33–41). Athens, Greece.
Aronson, A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceedings of the AMIA annual symposium* (pp. 17–21). Washington, DC.
Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 268*, 509–512.
Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the ACL workshop on intelligent scalable text summarization* (pp. 10–17). Madrid, Spain.
Barzilay, R., & Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd annual meeting of the association for computational linguistics* (pp. 141–148). Ann Arbor, Michigan.
Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 1–7.
Brooks, A. D., & Sulimanoff, I. (2002). Evidence-based oncology project. *Surgical Oncology Clinics of North America, 11*, 3–10.
Cohen, A., & Hersh, W. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6*, 57–71.
Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR), 22*, 457–479.
Farghlay, A., & Hedin, B. (2003). Domain analysis and representation. In *Handbook for language engineers* (pp. 21–58). Stanford, CA: CSLI Publications.
Fiszman, M., Rindflesch, T.C., & Kilicoglu, H. (2004). Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL workshop on computational lexical semantics* (pp. 76–83). Boston, MA.
Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the DARPA speech and natural language workshop* (pp. 233–237). Harriman, NY.
Gay, C., Kayaalp, M., & Aronson, A. (2005). Semi-automatic indexing of full text biomedical articles. In *Proceedings of the AMIA annual symposium* (pp. 271–275). Washington, DC.
Hovy, E. (2005). Automated text summarization. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 583–598). Oxford University Press.
Humphrey, S., Rogers, W., Kilicoglu, H., Demner-Fushman, D., & Rindflesch, T. (2006). Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology, 57*, 96–113.
Hunter, L., & Cohen, K. B. (2006). Biomedical language processing: Perspective whats beyond PubMed? *Molecular Cell, 21*, 589–594.
Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics, 24*, 1–40.
Joshi, M., Pedersen, T., & Maclin, R. (2005). A comparative study of support vector machines applied to the word sense disambiguation problem for the medical domain. In *Proceedings of the 2nd Indian conference on artificial intelligence (IICAI-05)* (pp. 3449–3468). Pune, India.
Lapata, M., & Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th international joint conference on artificial intelligence* (pp. 1085–1090). Edinburgh, Scotland.
Lau, A., & Coiera, E. (2008). Impact of web searching and social feedback on consumer decision making: A prospective online experiment. *Journal of Medical Internet Research, 10*, E2.
Lin, C.-Y. (2004a). Looking for a few good metrics: Automatic summarization evaluation – How many samples are enough? In *Proceedings of the 4th NTCIR workshop on research in information access technologies information retrieval, question answering and summarization*. Tokyo, Japan.
Lin, C.-Y. (2004b). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop on text summarization branches out* (pp. 74–81). Barcelona, Spain.
Liu, H., Teller, V., & Friedman, C. (2004). A multi-aspect comparison study of supervised word sense disambiguation. *Journal of the American Medical Informatics Association, 11*, 320–331.
Louis, A., & Nenkova, A. (2008). Automatic summary evaluation without human models. In *Proceedings of the 1st text analysis conference (TAC 2008)*. Gaithersburg, MD.
Mani, I. (1999). *Advances in automatic text summarization*. Cambridge, MA: The MIT Press.
Mani, I. (2001). *Automatic summarization*. Amsterdam: J. Benjamins Pub. Co.
McInnes, B., Pedersen, T., & Carlis, J. (2007). Using UMLS concept unique identifiers (CUIs) for word sense disambiguation in the biomedical domain. In *Proceedings of the annual symposium of the american medical informatics association* (pp. 533–537). Chicago, IL.
Mihalcea, R., & Tarau, P. (2004). TextRank – Bringing order into text. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2004)* (pp. 404–411). Barcelona, Spain.
Moens, M. F. (2000). *Automatic indexing and abstracting of document texts*. Kluwer Academic Publishers.
Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys, 41*, 1–69.
Navigli, R., & Lapata, M. (2007). Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007)* (pp. 1683–1688). Hyderabad, India.
Nelson, S., Powell, T., & Humphreys, B. (2002). The unified medical language system (UMLS) project. In A. Kent & C. M. Hall (Eds.), *Encyclopedia of library and information science*. Marcel Dekker, Inc.
Pitler, E., Louis, A., & Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 544–554). Uppsala, Sweden.
Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2008)* (pp. 186–195). Honolulu, Hawaii.
Plaza, L., Díaz, A., & Gervás, P. (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* (pp. 53–56). Manchester, UK.

Ponzetto, S.P., & Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1522–1531). Uppsala, Sweden.

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)* (pp. 87–92). Prague, Czech Republic.

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

Saggion, H., Torres-Moreno, J.-M., Cunha, I.D., & SanJuan, E. (2010). Multilingual summarization evaluation without human models. In *Proceedings of the 23rd international conference on computational linguistics (COLING 2010): Poster volume* (pp. 1059–1067). Beijing, China.

Sanderson, M. (1994). Word sense disambiguation and information retrieval. In *Proceedings of the 17th ACM SIGIR conference* (pp. 142–151). Dublin, Ireland.

Savova, G. K., Coden, A., Sominsky, I. L., Johnson, R., Ogren, P. V., de Groen, P. C., et al (2008). Word sense disambiguation across two domains: Biomedical literature and clinical notes. *Journal of Biomedical Informatics, 41*, 1088–1100.

Schuemie, M., Kors, J., & Mons, B. (2005). Word sense disambiguation in the biomedical domain: An overview. *Journal of Computational Biology, 12*(5), 554–565.

Schwartz, A., & Hearst, M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific symposium on biocomputing (PSB 2003)* (pp. 451–462). Lihue, Hawaii.

Shi, Z., Melli, G., Wang, Y., Liu, Y., Gu, B., & Kashani, M.M., et al. (2007). Question answering summarization of multiple biomedical documents. In *Proceedings of the Canadian conference on AI* (pp. 284–295). Montreal, Quebec.

Sinha, R., & Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE international conference on semantic computing (ICSC 2007)* (pp. 363–369). Irvine, CA, USA.

Stevenson, M., Guo, Y., Gaizauskas, R., & Martinez, D. (2008). Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics, 9*, S7.

Tsatsaronis, G., Vazirgiannis, M., & Androutsopoulos, I. (2007). Word sense disambiguation with spreading activation networks generated from thesauri. In *Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007)* (pp. 1725–1730). Hyderabad, India.

Vadlapudi, R., & Katragadda, R. (2010). Quantitative evaluation of grammaticality of summaries. In *Proceedings of the 11th conference on intelligent text processing and computational linguistics (CICLing)* (pp. 736–747). Iasi, Romania.

Weeber, M., Mork, J., & Aronson, A. (2001). Developing a test collection for biomedical word sense disambiguation. In *Proceedings of AMIA annual symposium* (pp. 746–750). Washington, DC.

Westbrook, J., Coiera, E., & Gosling, A. (2005). Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association, 12*, 315–321.

Yoo, I., Hu, X., & Song, I.-Y. (2007). A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. *BMC Bioinformatics, 8*, S4.

Zhou, L., Lin, C.-Y., Munteanu, D.S., & Hovy, E. (2006). ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the human language technology conference of the NAACL, main conference* (pp. 447–454). New York, NY.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., & Cohen, K. (2007). Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics, 8*, 358–375.

Please cite this article in press as: Plaza, L., et al. Resolving ambiguity in biomedical text to improve summarization. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.09.005