BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 22 2010, pages 2889–2896
doi:10.1093/bioinformatics/btq555

Data and text mining

Advance Access publication October 7, 2010

Graph-based Word Sense Disambiguation of biomedical documents

Eneko Agirre[1], Aitor Soroa[1] and Mark Stevenson[2,*]

[1]IXA NLP Group, University of the Basque Country, Donostia, Basque Country and [2]Department of Computer Science, Sheffield University, 211 Portobello, Sheffield S1 4DP, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Word Sense Disambiguation (WSD), automatically identifying the meaning of ambiguous words in context, is an important stage of text processing. This article presents a graph-based approach to WSD in the biomedical domain. The method is unsupervised and does not require any labeled training data. It makes use of knowledge from the Unified Medical Language System (UMLS) Metathesaurus which is represented as a graph. A state-of-the-art algorithm, Personalized PageRank, is used to perform WSD.

Results: When evaluated on the NLM-WSD dataset, the algorithm outperforms other methods that rely on the UMLS Metathesaurus alone.

Availability: The WSD system is open source licensed and available from http://ixa2.si.ehu.es/ukb/. The UMLS, MetaMap program and NLM-WSD corpus are available from the National Library of Medicine http://www.nlm.nih.gov/research/umls/, http://mmtx.nlm.nih.gov and http://wsd.nlm.nih.gov. Software to convert the NLM-WSD corpus into a format that can be used by our WSD system is available from http://www.dcs.shef.ac.uk/~marks/biomedical_wsd/ under open source license.

Contact: m.stevenson@dcs.shef.ac.uk

Received on July 5, 2010; revised on September 16, 2010; accepted on September 26, 2010

1 INTRODUCTION

The biomedical scientific literature is now so large that automated tools are necessary to access it effectively (Chapman and Cohen, 2009). However, this process is made difficult by the fact that terms in natural language can be ambiguous, i.e. may refer to more than one possible concept. For example, in the biomedical domain the word 'cold' is ambiguous and can mean (at least) 'common cold', 'cold temperature' or 'cold sensation'. *Word Sense Disambiguation* (WSD) systems aim to solve this problem by identifying the meanings of ambiguous words in context (Agirre and Edmonds, 2006; Navigli, 2009). For example, WSD would aim to identify that the meaning of *cold* in the sentence *The role of zinc in treating cold symptoms* is 'common cold'. This information could be used to improve literature searches, by ensuring that the documents returned only contain ambiguous terms when they are used in a meaning that is relevant to the search, and is also beneficial for other applications that are useful for biomedical researchers, such as automated indexing, information extraction and knowledge discovery (Aronson *et al.*, 2000; Friedman, 2000; MacMullen and Denn, 2005).

This article describes an approach to WSD in the biomedical domain that is based on graph-based algorithms. Since the approach is unsupervised, it does not require any labeled training data and relies on information from the Unified Medical Language System (UMLS) Metathesaurus (Humphreys *et al.*, 1998) instead. The UMLS Metathesaurus is converted into a graph to which the Personalized Page Rank algorithm (Agirre and Soroa, 2009) is applied to carry out WSD.

Section 2 describes previous work on WSD in the biomedical domain and the use of graph-based algorithm for WSD. Section 3 describes our approach to WSD in the biomedical domain using the UMLS Metathesaurus and Personalized Page Rank algorithm. The approach is evaluated against a standard dataset and the results were analysed in Section 4. These results are discussed in Section 5. The conclusions are found in Section 6.

2 RELATED WORK

The problem of WSD has been explored since the 1950s and is regarded as an important stage in text processing (Agirre and Edmonds, 2006; 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 (Schuemie *et al.*, 2005). The most popular approaches for WSD in biomedicine are based on supervised learning, for example Joshi *et al.* (2005); Liu *et al.* (2004); Savova *et al.* (2008). Although studies on domain-independent WSD have shown that supervised approaches outperform alternative ones, they require labeled training examples which may not be available and are expensive to create. This limitation means that most supervised approaches (including those mentioned above) can only disambiguate a small sample of words for which training data can be found, and this limits their usefulness in practise. In the context of biomedicine, Humphrey *et al.* (2006) avoided this problem by using Medline as training data and exploiting information it contains about the source of each abstract. This approach assigns Semantic Types from the UMLS Metathesaurus but is unable to distinguish between meanings with the same Semantic Type.

Unsupervised approaches do not require labeled training examples and often make use of knowledge bases, such as the UMLS Metathesaurus. McInnes (2008) describes such an approach that uses the UMLS to generate textual definitions for the possible

*To whom correspondence should be addressed.

meanings of ambiguous terms. WSD is carried out by comparing the context of the ambiguous term with the definitions of each possible sense and choosing the one with the most words in common. The approach is evaluated against 13 terms from the NLM-WSD corpus (see Section 4) and performance of 48.11% reported.

Graph-based methods have recently become widely used for domain-independent knowledge-based WSD (Agirre and Soroa, 2009; Navigli and Lapata, 2007; Sinha and Mihalcea, 2007; Tsatsaronis *et al.*, 2007). These methods represent the knowledge base as a graph which is then analysed 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.

This article presents an unsupervised knowledge-based WSD algorithm which is capable of disambiguating all words that are ambiguous in the UMLS Metathesaurus. Relations in the UMLS Metathesaurus are used to create a graph which is analysed using the Personalized PageRank algorithm to rank possible meanings of ambiguous words based on their structural importance in the graph and their relation to the words in context. This algorithm has previously been applied in a domain-independent setting, using WordNet as the knowledge base (Agirre and Soroa, 2009), and shown to outperform other, more ellaborate, graph-based algorithms (Navigli and Lapata, 2007; Sinha and Mihalcea, 2007; Tsatsaronis *et al.*, 2007).

3 GRAPH-BASED WSD

3.1 PageRank and Personalized PageRank

The PageRank algorithm (Brin and Page, 1998) is a method for ranking the vertices on a graph according to their relative structural importance. It was originally developed to rank World Wide Web pages based on the number of pages that link to them. Here, we describe it as an algorithm for generic graphs. The next sections describe how to use it for WSD using the UMLS.

PageRank uses a random walk model, where a *random surfer* starts a walk from an arbitrary node in the graph and, at each step, chooses an outgoing edge of the node at random to follows. The surfer may also decides to stop following edges and teleport to any node in the graph. The PageRank score of a vertex yields the probability that the random surfer is found in that vertex, assuming that the random walk continues indefinitely.

Specifically, let G be a graph with N vertices $(v_1,\ldots,v_N)$. For a given vertex v_i, let $\mathrm{In}(v_i)$ be the set of vertices pointing to it, and let d_j the out-degree of vertex v_j. The PageRank of vertex v_i is defined as:

$$P(v_i)=c\sum_{v_j\in \mathrm{In}(v_i)}\frac{1}{d_j}P(v_j)+(1-c)\frac{1}{N} \qquad (1)$$

where c is the so-called *damping factor*, a scalar value between 0 and 1. The PageRank for a vertex v_i is the addition of two terms. The first term models the probability of the random surfer arriving to v_i following the edges going from any vertex v_j to v_i, given by the sum of the probabilities of each vertex v_j having an edge to v_i times the weight of the edge, as given by the inverse of the degree of v_j. The second term represents the probability of the surfer randomly jumping to any node with equal probability. The damping factor c models the relative importance of each of the two terms.

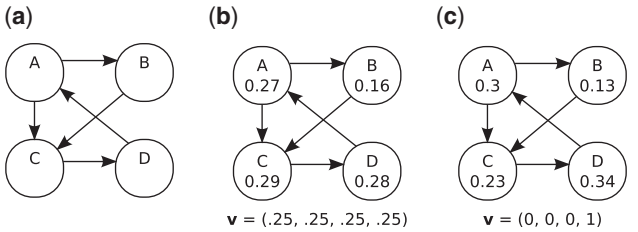

Fig. 1. Examples of a graph (**a**) alongside PageRank (**b**) and Personalized PageRank (**c**) computations for the graph. The number inside the circles are the PageRank value of the nodes.

The second term can also be seen as a smoothing factor that makes any graph fulfil the property of being aperiodic and irreducible, and thus guarantees that PageRank calculation converges to a unique stationary distribution.

PageRank is calculated by applying an iterative algorithm that computes Equation (1) repeatedly until convergence below a given threshold is achieved or until a pre-specified number iterations have been executed. The damping factor is usually set in the range [0.85..0.95]. Previous experiments (Agirre and Soroa, 2009) lead us to choose a damping factor of 0.85.

Figure 1 shows a sample graph (a) and the PageRank values for this graph (b). Initially, P for all four nodes are initialized with a uniform distribution, i.e. 0.25.[1] Given a damping factor of 0.85, in the first iteration the PageRank values are updated as follows:

$$\begin{aligned}
P(A^1)&=0.85\times & P(D^0)\times 1.0 & & +0.15\times 0.25=0.25\\
P(B^1)&=0.85\times & P(A^0)\times 0.5 & & +0.15\times 0.25=0.14\\
P(C^1)&=0.85\times & (P(A^0)\times 0.5+P(B^0)\times 1.0) & & +0.15\times 0.25=0.36\\
P(D^1)&=0.85\times & P(C^0)\times 1 & & +0.15\times 0.25=0.25
\end{aligned}$$

where the superscripts correspond to the current iteration, i.e. $P(A^0)$ corresponds to the initial value and $P(A^1)$ to the first iteration. The second iteration would calculate $P(A^2)$ based on $P(D^1)$, and so on. After a few iterations, convergence is attained and the PageRank values shown in graph (Fig. 1b) are obtained.

In certain situations, including graph-based WSD, we would like to include information about the relative importance of vertices in the graph. That is, given a set of vertices of interest, we would like to know which other vertices are closely related to them in the graph. For instance, we may be interested to know which nodes in graph (Fig. 1a) are closely related to node D (as shown in Fig. 1c).

Personalized PageRank (Haveliwala, 2002) computes the structural importance of the vertices in a graph when some vertices are more relevant than others for the task at hand. In order to introduce Personalized PageRank, we first rewrite Equation (1) in compact form by using matrices as follows. Let M be a $N\times N$ transition probability matrix, where $M_{ji}=\frac{1}{d_i}$ if a link from v_i to v_j exists, and zero otherwise. Let $\mathbf{v}$ be a stochastic normalized $N\times 1$ vector whose elements are all $\frac{1}{N}$. Then, the calculation of the *PageRank Vector* $\mathbf{P}$ over the graph G is equivalent to resolving the following Equation:

$$\mathbf{P}=cM\mathbf{P}+(1-c)\mathbf{v} \qquad (2)$$

[1]Note that the initial distribution values do no affect the final PageRank, provided that the algorithm converges.

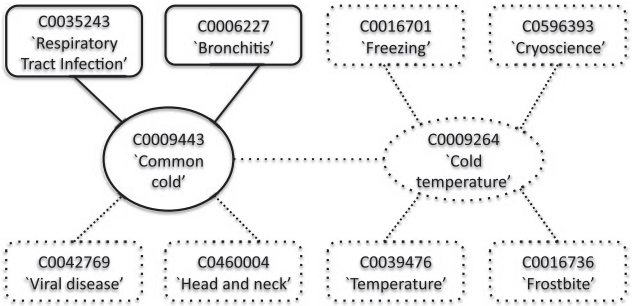

Fig. 2. Small part of the graph created from the UMLS MRREL table showing two possible concepts for the ambiguous word 'cold' (represented as ovals) and concepts connected to them (represented as rectangles). Concepts that would be assigned high weight after running personalized PageRank for the context 'To evaluate antibiotic-prescribing practices for children younger than 18 years who had received a diagnosis of **cold**, upper respiratory tract infection (URI), or bronchitis in the United States' are represented using solid lines, and the correct meaning of 'cold' (C0009443 'Common cold') would be selected for this example.

In PageRank, the vector **v** is uniformly distributed, thereby assigning equal probabilities to all vertices in the graph when random jumps are made. However, in *Personalized PageRank* the vector **v** can be non-uniform and assign stronger probabilities to certain vertices, effectively biasing the resulting PageRank vector to prefer these vertices. For example, if we concentrate all the probability mass on a unique vertex v_x, all random jumps on the walk will return to v_x and consequently its rank will be high; moreover, the high rank of v_x will cause all the vertices in its vicinity to also receive high rank. The importance of vertex v_x in the initial distribution of **v** then spreads through the graph during successive iterations of the algorithm. In this case, the personalized **P** vector represents the importance of every vertex in the graph relative to vertex v_x. Personalized PageRank can be solved using the same kind of algorithms as standard PageRank.

Figure 1 shows the result of standard PageRank in (b), where **v** is uniform, and the result of Personalized PageRank in (c) for the case where **v** is set to 0 for all vertices except *D*, which is set to 1. For standard PageRank, vertex C receives a PageRank value of 0.29 [as in graph (Fig. 1b)], meaning that a random surfer on this graph would spend 29% of the time on that node. *C* has the highest rank among the graph nodes, and therefore node *C* is the most important node in the graph. On the contrary, if we use Personalized PageRank and the random surfer makes all random jumps to *D* [as in graph (Fig. 1c)], then the rank of *D* is now the highest, followed by node *A*—which is linked directly to *D*— and nodes *C* and *B*.

3.2 Using Personalized PageRank for WSD

To use Personalized PageRank for WSD, the UMLS is represented as a graph in which the concepts are vertices and relations between them edges. Given the context of an ambiguous word (e.g. 'cold' in the context mentioned in Fig. 2), WSD is carried out by initializing **v** with equal values for all concepts that appear in the context (and zero for the rest), applying Personalized PageRank and then selecting the concept corresponding to 'cold' that has the highest PageRank.

Two sources of information are required for the Personalized PageRank algorithm to be used for WSD: a Knowledge Base and a dictionary. The Knowledge Base (KB) consists of a set of concepts and relations between them. It can be naturally represented as an undirected graph $G=(V,E)$ where nodes represent KB concepts (v_i) and the relation between concepts v_i and v_j is represented by an undirected edge $e_{i,j}$. The dictionary maps words and phrases found in documents to their possible concepts in the KB.

3.2.1 Knowledge Base The UMLS Metathesaurus is used as the KB. It is created by unifying a diverse range of controlled vocabularies and classification systems. The Metathesaurus is organized around concepts and each 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 also contains a wide range of information about the relations between CUIs in the form of database tables. The MRREL table lists relations between CUIs found in the various sources that comprise the Metathesaurus. This table lists a range of different types of connections between CUIs. For example, C0009443 'Common Cold' is related to C0035243 'Respiratory Tract Infections' by the PAR (parent) relation. Other types of relations in the MRREL table include QB (can be qualified by), RQ (related and possibly synonymous) and RO (related, other). For example, C0009443 'Common Cold' is related to C0460004 'Head and Neck' by the RO relation. Figure 2 shows a small part of the graph. The MRREL table also lists the source vocabulary from which the relation was obtained. For example, it states that the connection between C0009443 and C0460004 is found in the National Cancer Institute Thesaurus. The same relation may also be found in multiple source vocabularies, e.g. the CUIs C0009443 and C0035243 are related in four separate vocabularies.

Co-occurrence relations between CUIs are found in the MRCOC table. These relations exist between similar concepts (e.g. C00004238 'Atrial Fibrillation' and C0003811 'Cardiac Arrhythmia') or different concepts that share an important connection (e.g. C00004238 'Atrial Fibrillation' and C0012265 'Digoxin'). Although the MRCOC table lists a large number of co-occurrence relations, most concepts do not have any co-occurrence relations associated with them. The information in this table was created by automatically processing three information sources (Medline, 2002–2007; AI/RHEUM, 1993 and the Canonical Clinical Problem Statement System, 1999) to identify co-occurrences. The MRCOC table includes details about the strength of the co-occurrence relation between concepts based on the number of co-occurrences identified.

It is straightforward to convert the information contained in the MRREL and MRCOC tables into a graph. The concepts form the vertices with the relations listed in the tables being used to define the edges between them. No weights are used for the relations that are extracted from the MRREL table. In the case of the MRCOC table, we did use the strength of co-occurrence to produce some subsets of the graph (see Section 4.2).

3.2.2 Dictionary The dictionary contains mappings from words and phrases in text to UMLS CUIs. It is created using the MetaMap program (Aronson, 2001) that splits the input text into phrases and maps each onto the set of possible CUIs that they could refer to, known as *candidates*. The set of candidates for each word or phrase in the context of the ambiguous terms are extracted from

the MetaMap output and used to create the dictionary to define the possible CUIs for each word in its context.

3.2.3 Static PageRank baseline Applying the traditional PageRank algorithm over the graph created from the UMLS leads to all CUIs being ranked according to their PageRank value, i.e. a context-independent ranking of CUIs. This can be used to create a WSD system by examining the relative rankings of the CUIs for a target word and returning the highest ranking one. We call this application of PageRank to WSD *Static PageRank*, since it does not change with the context, and use it as a baseline. The *static* baseline favours concepts with high degree, and thus disambiguates each word to the concept having most connections in the graph, regardless of context.

3.2.4 Personalized PageRank Static PageRank is independent of context, but this is not what we want in a WSD system. Given an input piece of text we want to disambiguate all content words (i.e. nouns, verbs, adjectives and adverbs) in the input based on the words in the context. This can be achieved using Personalized PageRank as follows.

Given an input text, we extract the list $W = \{W_1, \ldots, W_m\}$ of content words which have an entry in the dictionary and can therefore be related to UMLS concepts. Note that monosemous words will be attached to just one concept, whereas polysemous words may be attached to several. The context words are first inserted into G as nodes, and linked with directed edges to their respective concepts. The Personalized PageRank of the graph G is then computed by concentrating the initial probability mass uniformly over the newly introduced word nodes. As the words are linked to the concepts by directed edges, they act as source nodes injecting mass into the concepts they are associated with, which thus become relevant nodes and spread their mass over the UMLS graph. The resulting Personalized PageRank vector can be seen as a measure of the relevance of UMLS concepts given the context. As a result of the disambiguation process, every UMLS concept receives a score. Each target word can then be disambiguated by examining each of its possible concepts in the graph, G, and selecting the one with the highest score. Figure 2 shows an example.

A problem occurs if the possible CUIs of the target word being disambiguated are themselves related. In this situation, those CUIs reinforce each other and reduce the influence of the other senses in the context. With this observation in mind, we introduce a change in the algorithm: for each target word W_i, we concentrate the initial probability mass in the senses of the words surrounding W_i, but not in the senses of the target word itself, so that context words increase its relative importance in the graph. The main idea of this approach is to avoid biasing the initial score of concepts associated to target word W_i, and let the surrounding words decide which concept associated with W_i has more relevance.

3.2.5 Interoperability and performance of the system UKB is open source, programmed in C++ and easily integrated in third-party software as a library. For instance, the open source multilingual text-processing package Freeling[2] incorporates UKB.

The steps needed to run our system are as follows. Before performing WSD, the `MRREL` and `MRCOC` tables from the UMLS need to be converted to a binary graph format. Given a target document, we first run MetaMap to construct the dictionary for the WSD system. The Personalized PageRank algorithm can then be run. This uses MetaMap's output for the target document, the graph and the dictionary. It outputs the disambiguated concepts in the form of CUI numbers with weights.

The performance of our system on a PC with 2 QuadCore Xeon processors at 3160 MHz and 32 G of memory was the following: building the binary graph from the UMLS tables takes 21.8 s, loading the binary graph takes 5.6 s and 1.6 G of memory. WSD is performed at a rate of 37 instances per minute.

[2] `http://nlp.lsi.upc.edu/freeling/`

4 EVALUATION

The WSD system was evaluated using the NLM-WSD corpus Weeber *et al.* (2001). This contains 50 ambiguous terms with 100 instances of each. The instances are abstracts containing the ambiguous term randomly extracted from those added to Medline in 1998. The 5000 instances were manually disambiguated by 11 annotators, who tagged each occurrence of the target term with the corresponding meaning. Some instances were tagged 'None' to indicate that the annotators did not consider any of the possible meanings in UMLS applied. Following standard practice (Humphrey *et al.*, 2006; McInnes, 2008), these instances were not used in the evaluation, yielding a total of 3983 examples and 49 terms. (One term, 'association', was excluded since all 100 instances were labeled as 'None'.)

In addition to the full NLM-WSD dataset, a subset of 13 of these terms was also used for evaluation. This subset was used by McInnes (2008) and consists of terms that have a majority sense that accounts for less than 65% of the instances and whose possible senses do not share the same semantic type.

A window of 20 terms around the target word (i.e. the 10 preceding and 10 following terms) are used as the context. This is created by using MetaMap to identify the terms around the target word (see Section 3.2.2). Any phrases that are not mapped onto a CUI are discarded and the terms that form each of the remaining phrases are used to create the context. The damping factor (Section 3.1) was set to 0.85. The values of these parameters were selected based on previous work (Agirre and Soroa, 2009). Section 4.4 reports a *post hoc* analysis exploring the effects of varying them.

The 2007AB version of the UMLS was used for the experiments. This version was chosen since we had access to a mapping between the NLM-WSD sense labels and UMLS CUIs. Such a mapping is only required to evaluate our approach and it would be possible to use the approach described in this article to carry out WSD relative to any version of the UMLS. The mapping from the 2007AB version of the UMLS was created with the assistance of publicly available software and manually verified.

4.1 Results

Table 1 shows results of the system evaluation. Performance is measured by accuracy, the percentage of instances correctly disambiguated. Note that our algorithm returns a sense for all instances. The confidence interval, computed using bootstrap resampling with 95% confidence (Noreen, 1989), is also shown.

The top part of the table shows the results on the full NLM-WSD dataset. The first two rows show the result using Personalized

Table 1. Main results over the NLM-WSD dataset and the subset used in McInnes (2008)

Method	KB	Acc.
Full NLM-WSD Dataset		
ppr	MRREL	**68.1** [66.80, 69.23]
ppr	MRREL+MRCOC	65.5 [64.30, 66.73]
Static	MRREL	58.4 [57.07, 59.60]
Random	–	45.6
McInnes (2008) subset of NLM-WSD		
ppr	MRREL	**55.0** [52.60, 57.76]
McInnes (2008)	-	48.1

Our ppr method (*ppr*) is shown together with the *static* and *random* baselines, as well as the results by McIness.

PageRank (labeled *ppr*) using the graph created from the MRREL table alone and the combination of the MRREL and MRCOC tables. (We were not able to use the MRCOC table alone since it does not include all of the CUIs in the NLM-WSD corpus.) Performance using two baselines are also included. The first, *static*, applies the Personalized PageRank algorithm to the graph without making use of context (see Section 3.2.3) and the second, *random*, simply chooses a meaning at random. All differences in the table are statistically significant. Personalized PageRank clearly outperforms both the *random* and *static* baselines.

The best performance is obtained using the graph created from the MRREL table. Results decrease when the extra relations from the MRCOC table are added. This drop in performance was unexpected since co-occurrence information is generally considered to be very useful for WSD (Agirre and Edmonds, 2006). However, the MRCOC table only contains relations for some CUIs, unlike the MRREL table which contains relations for all CUIs. This negatively affects Personalized PageRank since it is more likely to select CUIs that are more highly connected and the fact that some CUIs do not appear in the MRCOC table creates a bias towards those which do. Analysis indicated that there are only four terms ('ganglion', 'man', 'secretion' and 'surgery') for which all of the possible CUIs appear in the MRCOC table. For 25 terms some of the CUIs appear in the MRCOC table while others do not. In addition, relations in the MRCOC table are generated automatically and may also be noisy.

The lower part of Table 1 reports results on the 13 terms used by McInnes (2008). The best approach, Personalized Pagerank using MRREL, achieves better results than those reported by McInnes (2008), the state-of-the-art in knowledge-based WSD. Possible reasons for this improved performance are that the MRREL table contains information that is more useful for WSD than the CUI definitions used by McInnes (2008) and that the graph-based algorithm used in our approach benefits from being able to make use of information from the entire UMLS.

4.2 UMLS subsets

In this section, we explore the effect of using subsets of the UMLS: relations from various vocabularies in MRREL and ranked relations from MRCOC.

A greedy algorithm (Chvatal, 1979) was used to identify a set of vocabularies that included the CUIs used as possible senses of the terms in the NLM-WSD dataset. One vocabulary, MTH (UMLS Metathesaurus), was excluded from the set of vocabularies considered since it consists of concepts that were created specifically to create the Metathesaurus, rather than being an independent vocabulary in its own right. The greedy algorithm generated a set of four vocabularies: AOD (Alcohol and Other Drug Thesaurus), MSH (Medical Subject Headings), CSP (Crisp Thesaurus) and SNOMEDCT (SNOMED Clinical Terms). One of the possible senses for 'resistance' is only found in the MTH vocabulary and this term cannot be represented using these vocabularies. Note that the union of all four subsets covers all senses of the target words, but does not contain all relations in MRREL.

Table 2. Results using different vocabularies of the MRREL table

KB	#CUIs	#relations	Acc.	Terms	MRREL
AOD	15 901	58 998	51.5	4	61.5
MSH	278 297	1 098 547	44.7	9	66.6
CSP	16 703	73 200	60.2	3	69.4
SNOMEDCT	304 443	1 237 571	62.5	29	68.8
All above	572 105	2 433 324	64.4	48	68.0

The table shows the number of CUIs and relations in the graph (#CUIs and #relations), WSD accuracy using the graph with the Personalized PageRank algorithm (Acc.), number of terms to which the graph can be applied (Terms) and the results on those terms using the entire MRREL table (MRREL).

Table 3. WSD results using different subsets of the MRCOC table

KB	#relations	Acc.
MRREL	5 352 190	**68.1**
MRREL+MRCOC$_1$	6 096 974	68.0
MRREL+MRCOC$_2$	7 546 138	66.9
MRREL+MRCOC$_{full}$	11 362 335	66.0

The table shows the number of relations in the graph (#relations) and WSD accuracy over all terms (Acc).

Table 2 shows the results when the Personalized PageRank algorithm is applied to graphs created using single vocabularies and the combination of all four. No single vocabulary includes all possible concepts for every sense and the column marked 'Terms' indicates the number of terms for which all possible concepts are included in a vocabulary or set of vocabularies. Results in the 'Acc.' column list the WSD performance over those terms using the graphs created using the single vocabularies (or their combination) while the 'MRREL' column lists the results using the graph created from the MRREL table over the same terms. Performance using individual vocabularies, or the combination of four vocabularies, is always lower than when the full MRREL graph is used. This indicates that our algorithm is able to exploit information from the multiple vocabularies that are combined to form the MRREL table in the UMLS and that including additional vocabularies, even ones that are not necessary to represent all of the possible meanings for ambiguous terms, improves WSD performance.

Table 3 reports results when adding several subsets of MRCOC to the MRREL relations. Instead of using all relations in MRCOC, we aim to identify the most useful ones using the Mutual Information (MI) statistic (Church and Hanks, 1990) that ranks pairs of concepts based on the probability that they occur more frequently than would

be expected by chance. Several subsets of MRCOC relations were generated by ranking them by MI and successively adding those with the highest score to the graph created from the MRREL table. MRREL + $MRCOC_1$ adds approximately 750 000 new co-occurrences and MRREL + $MRCOC_2$ adds around 2 million.

The table shows that performance drops systematically as co-occurrence relations from the MRCOC table are added to the graph. This situation is somewhat different from the positive effect of adding information from the MRREL table. These results enforce our hypothesis that co-occurrence relations negatively change the topology of the graph, degrading the performance of our algorithm. The smallest drop in performance is obtained when $MRCOC_1$ is added, suggesting that MI is a useful technique for selecting the most informative co-occurrence relations.

4.3 Word by word analysis

Table 4 lists the results for each of the individual terms using the MRREL graph. The columns labeled '#CUI' and '#inst' list the number of possible CUIs and instances for each term. The 'Single' column indicates the performance when a single UMLS vocabulary is used to create the graph, using one of the four subsets listed in Section 4.2. Missing figures indicate that the possible CUIs for a particular term are not included in any single vocabulary.[3] For some terms, all possible CUIs are included in multiple vocabularies and in these cases the best result is listed.[4] The column 'Subset' shows the results when a graph is created using the combination of the four vocabularies that were considered. (There is no value for 'resistance' since one of its senses is not included in the subset of four vocabularies, see Section 4.2.) Finally, the column 'Full' shows results when the graph created using the entire MRREL table from the UMLS is used.

Table 4 shows a significant variation in performance for individual terms with results ranging between 99% ('secretion') and 11.1% ('fit'). The PPR algorithm has a bias towards senses that are highly connected within the graph. For some terms there are significant differences between the connectivity for the possible senses. For example, one possible meaning of 'fit', C0036572 'Seizures', is linked to 1561 other CUIs in the MRREL table while the alternative meaning, C0424576 'Fit and well', is only linked to 18. The majority of errors for this term were caused by PPR assigning C0036572 to examples for which the correct CUI was C0424576.

The table also shows that the graph producing the best performance varies for each word. Overall, the graph created from the full MRREL table produces the best score for slightly more of the ambiguous terms (25) than either the individual vocabularies (20) or their combination (24). However, note that the overall average performance using the graph created from the full MRREL table is significantly better than when the subset is used (see Table 2).

4.4 Exploring context length and damping factor

Our algorithm has two free parameters: context length and damping factor of the PageRank formula. Default values for these parameters

[3]For example, the term 'adjustment' contains CUIs that are included in MSH and SNOMEDCT. However, neither vocabulary includes all three possible CUIs for this term.

[4]For example, the two CUIs for 'immunosuppression' are included in three vocabularies (CSP, MSH and SNOMEDCT) with SNOMEDCT producing the highest performance.

Table 4. Word by word analysis of WSD results

Word	#CUI	#inst	Single	Subset	Full
Adjustment	3	93		33.3	**35.5**
Blood pressure	3	100	**53.0**[d]	50.0	48.0
Cold	5	95	**32.6**[d]	26.3	28.4
Condition	2	92	**95.7**[d]	39.1	48.9
Culture	2	100		33.0	**77.0**
Degree	2	65		**95.4**	93.8
Depression	2	85	16.9[d]	64.3	**94.1**
Determination	2	79		49.4	**94.9**
Discharge	2	75	24.0[d]	61.3	**69.3**
Energy	2	100		**73.0**	27.6
Evaluation	2	100	**59.0**[d]	54.0	50.0
Extraction	2	87		23.0	**27.6**
Failure	2	29		27.6	**72.4**
Fat	2	73	56.2[d]	63.0	**95.9**
Fit	2	18	**16.7**[d]	11.1	11.1
Fluid	2	100	83.0[d]	**92.0**	**92.0**
Frequency	2	94	**98.9**[d]	**98.9**	**98.9**
Ganglion	2	100	66.0[c]	**77.0**	64.0
Glucose	2	100	**91.0**[d]	**91.0**	90.0
Growth	2	100	**37.0**[c]	**37.0**	**37.0**
Immunosuppression	2	100	**64.0**[d]	59.0	62.0
Implantation	2	98	75.0[b]	**84.7**	**84.7**
Inhibition	2	99		**24.2**	22.2
Japanese	2	79	**70.9**[d]	**70.9**	64.6
Lead	2	29	**93.1**[d]	**93.1**	**93.1**
Man	2	92	**61.5**[a]	34.8	44.6
Mole	3	84	10.7[d]	**63.1**	27.4
Mosaic	3	97		60.8	**66.0**
Nutrition	3	89		**33.7**	32.6
Pathology	2	99		**34.3**	28.3
Pressure	3	96	52.1[d]	69.8	**97.9**
Radiation	2	98	**58.2**[d]	53.1	53.1
Reduction	2	11	36.4[d]	**54.5**	**54.5**
Repair	2	68	63.2[d]	72.1	**76.5**
Resistance	2	3			66.7
Scale	3	65	67.7[d]	52.3	**84.6**
Secretion	2	100	**99.0**[c]	**99.0**	**99.0**
Sensitivity	3	51		**41.2**	27.5
Sex	3	100	84.0[d]	**85.0**	**85.0**
Single	2	100		80.0	**82.0**
Strains	2	93	92.5[d]	**97.8**	96.8
Support	2	10	**80.0**[d]	**80.0**	**80.0**
Surgery	2	100	95.9[c]	**97.0**	**97.0**
Transient	2	100		97.0	**99.0**
Transport	2	94	**98.9**[d]	**98.9**	69.1
Ultrasound	2	100	**84.0**[c]	**84.0**	83.0
Variation	2	100	**85.0**[d]	80.0	75.0
Weight	2	53	**56.6**[d]	**56.6**	**56.6**
White	2	90	**68.9**[a]	67.8	63.3
Best result			20	24	25

The 13 terms used by McInnes (2008) are printed underlined. In the column labeled 'Single', the following symbols are used to indicate the vocabulary that is used to create the graph that was used to produce the result: [a] AOD, [b] CSP, [c] MSH and [d] SNOMEDCT. Where more than one single vocabulary contains all possible meanings for a particular term, results are shown for the vocabulary that produced the best performance. The best result for each term is printed in **bold font**. The bottom row, *Best result*, shows the number of terms for which the best result is found in that column. Missing figures indicate that the possible senses of the ambiguous term are not included in the graph.

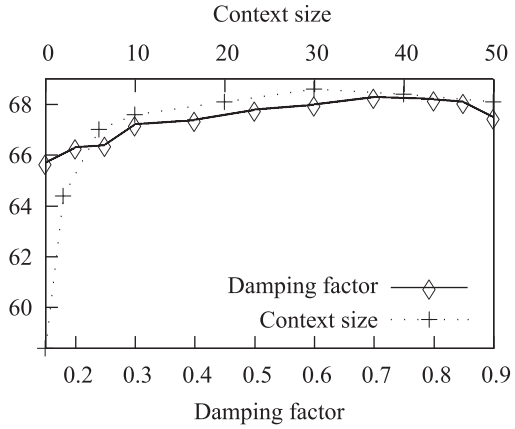

Fig. 3. Accuracy results when exploring different values for context size (upper *x*-axis, crosses) and damping factor (lower axis, diamonds) on the NLM-WSD dataset.

based on previous work were used for the results reported so far: a context length of 20 terms and damping factor of 0.85. Figure 3 shows the results obtained using different values for these two parameters.

Figure 3 shows that the best results are obtained using a context of 30 words. However, the difference in performance compared to using our default context (20 words) is relatively small. Performance deteriorates when the context is limited (shorter than 10 terms). The lowest performance is obtained when no context is used (i.e. context size of 0), which corresponds to the *static* algorithm. Our algorithm is robust to the actual context length used, given a minimum amount of context.

The best results when the damping factor was varied were obtained using a damping factor of 0.70, although performance was very close to the default value used in our experiment (0.85). This shows that the method is robust to changes in damping factor, provided it does not vary too far from the values suggested in the literature (Haveliwala, 2002).

5 DISCUSSION

There is some debate on how accurate WSD performance has to be to assisted in applications. Sanderson (1994) carried out experiments suggesting that a WSD system would need to correctly disambiguate 90% of words in order to be useful for Information Retrieval. However, more recent experiments (Agirre *et al.*, 2009; Caputo *et al.*, 2009) have shown that WSD with lower performance can improve Information Retrieval results, showing that the way in which the output of the WSD system is used is as important as the WSD performance. It has also been shown that WSD can improve several applications including Cross-lingual Information Retrieval (Stevenson and Clough, 2004), Machine Translation (Chan *et al.*, 2007) and Information Extraction (Chai and Biermann, 1999; Surdeanu *et al.*, 2008). Performance of the WSD components of these systems is generally not reported but it is likely that it will be lower than the result obtained by the best system in a community evaluation exercise of all-words WSD systems, 65% (Snyder and Palmer, 2004). This suggests that the WSD accuracy obtained by our system could be used to improve performance of applications.

The results reported here are based on the only available dataset, NLM-WSD, which contains terms that are frequent and ambiguous (Weeber *et al.*, 2001). The inter-annotator agreement for this data set is relatively low [kappa score 0.47 (Savova *et al.*, 2008)] which indicates that disambiguating these examples is not easy for humans. The performance of our approach for all ambiguous terms in a document cannot be reliably extrapolated but may be higher than the results reported here given the challenging nature of the NLM-WSD dataset.

It is possible that the performance of our algorithm could be further improved by adding other parts of the UMLS, such as the Semantic Types, to the graph, or by making use of domain information from the MeSH codes, which has been shown to be useful for supervised WSD in the biomedical domain (Stevenson *et al.*, 2008).

6 CONCLUSIONS

This article presents a WSD system for biomedical documents. The system is unsupervised and is able to disambiguate all words that are ambiguous in the UMLS Metathesaurus. Disambiguation is carried out by converting tables from the UMLS Metathesaurus into a graph and using the Personalized PageRank algorithm to select the best sense for each ambiguous word. Experiments show that the best results were obtained using the combination of all vocabularies in the `MRREL` table of the Metathesaurus. Performance of the approach reported here surpasses results reported for other systems that used the UMLS Metathesaurus as a knowledge source.

ACKNOWLEDGEMENTS

We are grateful to Bridget McInnes for providing the mapping between the 2007AB version of the UMLS and the NLM-WSD sense labels.

Funding: Engineering and Physical Sciences Research Council (EP/D069548/1); Ministry of Science and Innovation (KNOW2 project, TIN2009-14715-C04-01).

Conflict of Interest: none declared.

REFERENCES

Agirre,E. and Edmonds,P. (eds) (2006). *Word Sense Disambiguation: Algorithms and applications*. Springer, Berlin.

Agirre,E. and Soroa,A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, pp. 33–41.

Agirre,E. *et al.* (2009). CLEF 2009 ad hoc track overview: robust - WSD task. In *Working Notes of the Cross-Lingual Evaluation Forum*. Corfu, Greece.

Aronson,A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the American Medical Informatics Association (AMIA)*, pp. 17–21.

Aronson,A. *et al.* (2000). The NLM indexing initiative. In *Proceedings of the AMIA Symposium* . Los Angeles, CA, pp. 17–21.

Brin,S. and Page,L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Networks ISDN Systems*, **30**, 107–117.

Caputo,A. *et al.* (2009). From fusion to re-ranking: a semantic approach. In *Proceeding of the 33rd International ACM SIGIR Conference.*Geneva, Switzerland, pp. 815–816.

Chai,J. and Biermann,A. (1999). The use of word sense disambiguation in an information extraction system. In *Proceedings of the Eleventh Annual Conference on Innovative Applications of Artificial Intelligence*. Portland, OR, pp. 850–855 .

Chan,Y.S. *et al.* (2007). Word Sense Disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pp. 33–40.

Chapman,W. and Cohen,K. (2009). Current issues in biomedical text mining and natural language processing. *J. Biomed. Informat.*, **42**, 757–759.

Church,K. and Hanks,P. (1990). Word association norms, mutual information, and lexicography. *Computat. Linguistics*, **16**, 22–29.

Chvatal, V. (1979). A greedy heuristic for the set-covering problem. *Math.Operat. Res.*, **4**, 233–235.

Friedman,C. (2000). A broad coverage natural language processing system. In *Proceedings of the AMIA Symposium*. Los Angeles, CA, pp. 270–274.

Haveliwala,T.H. (2002). Topic-sensitive PageRank. In *WWW '02: Proceedings of the 11th International Conference on World Wide Web*. ACM, New York, NY, USA, pp. 517–526 .

Humphreys,L. *et al.* (1998). The Unified Medical Language System: an Informatics Research Collaboration. *J. Am. Med. Informat. Assoc.*, **1**, 1–11.

Humphrey,S. *et al.* (2006). Word Sense Disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. *J. Am. Soc. Informat. Sci. Technol.*, **57**, 96–113.

Joshi,M. *et al.* (2005). A comparative study of support vector machines applied to the Word Sense Disambiguation problem for the Medical Domain. In *Proceedings of the Second Indian Conference on Artificial Intelligence (IICAI-05)*. Pune, India, pp. 3449–3468, .

Liu,H. *et al.* (2004). A multi-aspect comparison study of supervised Word Sense Disambiguation. *J. Am. Med. Informat. Assoc.*, **11**, 320–331.

MacMullen,J. and Denn,O. (2005). Information problems in Molecular Biology. *J. Am. Soc. Informat. Sci. Technol.*, **56**, 447–456.

McInnes,B. (2008). An unsupervised vector approach to biomedical term disambiguation: integrating UMLS and Medline. In *Proceedings of the ACL-08: HLT Student Research Workshop* . Association for Computational Linguistics, Columbus, Ohio, pp. 49–54.

Navigli,R. (2009). Word Sense Disambiguation: a survey. *ACM Comput. Surv.*, **41**, 1–69.

Navigli,R. and Lapata,M. (2007). Graph connectivity measures for unsupervised word sense disambiguation. In *Proceeding of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*. Hyderabad, India, pp. 1683–1688.

Noreen,E.W. (1989). *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons, New York.

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

Savova,G.K. *et al.* (2008). Word sense disambiguation across two domains: biomedical literature and clinical notes. *J. Biomed. Informat.*, **41**, 1088–1100.

Schuemie,M. *et al.* (2005). Word Sense Disambiguation in the Biomedical Domain: an overview. *J. Comput. Biol.*, **12, 5**, 554–565.

Sinha,R. and 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)*. Irvine, CA, USA.

Snyder,B. and Palmer,M. (2004). The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain, pp. 41–43.

Stevenson,M. and Clough,P. (2004). EuroWordNet as a resource for cross-language Information Retrieval. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal, pp. 777–780.

Stevenson,M. *et al.* (2008). Disambiguation of biomedical text using a variety of knolwedge sources. *BMC Bioinformatics*, **9**(Suppl. 11), S7.

Surdeanu,M. *et al.* (2008). Learning to rank answers on large online QA collections. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, pp. 719–727.

Tsatsaronis,G. *et al.* (2007). Word sense disambiguation with spreading activation networks generated from thesauri. In *IJCAI*, pp. 1725–1730.

Weeber,M. *et al.* (2001). Developing a test collection for Biomedical Word Sense Disambiguation. In *Proceedings of the AMIA 2001 Symposium*, pp. 46–750.