
Article:
Duque, A., Stevenson, R.M. orcid.org/0000-0002-9483-6006, Martinez-Romo, J. et al. (1 more author) (2018) Co-occurrence graphs for word sense disambiguation in the biomedical domain. Artificial Intelligence in Medicine, 87. pp. 9-19. ISSN 0933-3657

https://doi.org/10.1016/j.artmed.2018.03.002

Co-Occurrence Graphs for Word Sense Disambiguation in the Biomedical Domain

Andres Duque[a,*], Mark Stevenson[b], Juan Martinez-Romo[a], Lourdes Araujo[a]

[a]*NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos. Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain*
[b]*Department of Computer Science, University of Sheffield. Regent Court, 211 Portobello. S1 4DP Sheffield, United Kingdom*

Abstract

Word Sense Disambiguation is a key step for many Natural Language Processing tasks (e.g. summarization, text classification, relation extraction) and presents a challenge to any system that aims to process documents from the biomedical domain. In this paper, we present a new graph-based unsupervised technique to address this problem. The knowledge base used in this work is a graph built with co-occurrence information from medical concepts found in scientific abstracts, and hence adapted to the specific domain. Unlike other unsupervised approaches based on static graphs such as UMLS, in this work the knowledge base takes the context of the ambiguous terms into account. Abstracts downloaded from PubMed are used for building the graph and disambiguation is performed using the Personalized PageRank algorithm. Evaluation is carried out over two test datasets widely explored in the literature. Different parameters of the system are also evaluated to test robustness and scalability. Results show that the system is able to outperform state-of-the-art knowledge-based systems, obtaining more than 10% of accuracy improvement in some cases, while only requiring minimal external resources.

Keywords: Word Sense Disambiguation, Graph-Based Systems, Unsupervised Machine Learning, Unified Medical Language System, Natural Language Processing, Information Extraction.

1. Introduction

The vast amount of unstructured textual information available in the biomedical sciences has created the need for automatic systems to access, retrieve and process these documents [1]. However, this is made more difficult by the range of lexical ambiguities they contain, including different meanings of general terms or the different extended forms of acronyms and abbreviations. For example, the word "surgery" may refer to the branch of medicine that applies operative procedures

*Corresponding author
Email addresses: aduque@lsi.uned.es (Andres Duque), mark.stevenson@sheffield.ac.uk (Mark Stevenson), juaner@lsi.uned.es (Juan Martinez-Romo), lurdes@lsi.uned.es (Lourdes Araujo)

to treat diseases, or to one of those operative procedures. Also, the acronym "BSA" could refer to multiple expansions such as "Bovine Serum Albuminum" and "Body Surface Area". There exist many different types of lexical ambiguity in biomedical documents, which represents an additional challenge when performing WSD in this domain [2]: words and phrases with more than one possible meaning, abbreviations with more than one possible expansion, or names of genes which may also contain ambiguity when standard naming conventions are not followed (the names of more than one thousand gene terms are standard English words [3]).

In this work, we present an unsupervised technique for addressing the Word Sense Disambiguation (WSD) problem in the biomedical domain. This technique, based on the mathematical background developed in [4], relies on the creation of a co-occurrence graph from a set of documents. This graph represents relations between pairs of words or concepts that appear frequently in the same document.

The contributions of this paper are to introduce a novel graph-based approach for WSD in the biomedical domain and, by evaluating it using datasets containing a range of ambiguities, demonstrate that it outperforms alternative approaches that do not make use of external knowledge sources.

The rest of the paper is organised as follows. Section 2 provides background on different approaches to biomedical WSD found in the literature. Section 3 describes the proposed system, detailing the different steps involved in the disambiguation process. Evaluation is carried out using two datasets (see Section 4) with the results described in Section 5. Finally, conclusions and future work are found in Section 6.

2. Previous Work

Regardless of whether we refer to general or specific domains, such as the biomedical one, it is commonly accepted in the literature [5, 6, 1] that most WSD algorithms fall into one of the following categories: techniques that need labelled training data, and knowledge-based techniques. The first category, also called supervised techniques, usually applies machine learning (ML) algorithms to labelled data to develop a model, based on features extracted from the context of the ambiguous words. The development of these features requires a comprehensive understanding of the problem being addressed [7]. We can find many different studies which address general WSD under this supervised point of view, through the use of classical machine learning algorithms [8], and in the last few years also adapting new techniques such as word embeddings [9]. When it comes to the biomedical domain, many works also belong to this category, making use of different ML approaches to address the problem [10, 11, 12, 13], although the bottleneck caused by the scarcity of labelled resources remains a major problem. Other semi-supervised works attempt to relieve this issue by introducing "pseudo-data" to the training examples [14, 15].

Knowledge-based methods use external resources as sources of information for performing WSD. As it happens with supervised methods, general WSD have been also addressed under this point of view. In particular, graph-based techniques using WordNet [16] as main knowledge base have been proved to present successful results in this kind of tasks [17, 18]. The dominant knowledge source in the biomedical domain is the Unified Medical Language System (UMLS) Metathesaurus [19], which assigns a Concept Unique Identifier (CUI) to each medical concept. These

concepts are then linked to other CUIs depending on the different relations between them [20]. Some methods directly convert this database into a graph [21], and use this graph for performing the disambiguation. Other works directly use information from the UMLS database for extracting additional information: In [22] second-order vectors are created by extracting textual information about each of the possible senses of an ambiguous term from UMLS. The method introduced in [23] makes use of information from the UMLS database through a statistical analysis. In this work, the knowledge base is used for calculating the probability $P(w_j|c_i)$, of finding a word w_j in any of the lexical forms related to a concept c_i, or to concepts linked to it in the database. Once that these probabilities have been found, the most suitable CUI related to an ambiguous term found in a context (tipically, the abstract of a biomedical paper, as we will observe in the definition of the test datasets) can be determined. For performing this disambiguation, the authors apply a method similar to Naïve Bayes which makes use of the words in the contexts, and those word-concept probabilities previously calcualted, for ranking the candidate CUIs for the ambiguous terms. Although this work presents some similarities to our system (for example, the statistical treatment of co-occurrences), the source of knowledge used for disambiguation is directly the UMLS database, while in our case, we built our own knowledge base in an unsupervised way from a corpus of biomedical documents.

Hence, and as we will explain later in more detail, the structured knowledge source that we use in the disambiguation phase of our method (the co-occurrence graph) is built automatically, exploiting the UMLS database to convert text from the original document set to medical concepts. However, this step can be seen as independent from the disambiguation process itself. We do not make use of any other external structured source of information in subsequent steps since the graph in which the disambiguation algorithm relies is directly built from those documents containing medical concepts. We will compare the results obtained by our system with other state-of-the-art knowledge-based systems addressing the same problem.

3. System Description

The co-occurrence graph used by the approach presented here is based on the hypothesis that documents are consistent, i.e., there is a strong tendency for the concepts found in a document to be related. Since this may not be true for all the concepts in the document, statistical analysis is applied to identify those concepts in documents that do not fulfill this hypothesis. In this analysis, only those pairs of concepts frequently co-occurring in the same documents are linked in the graph. This technique for building the co-occurrence graph has been previously used for general WSD tasks, such as Cross-Lingual WSD [24], with successful results, which suggests that a similar approach could also lead to competitive results in domain-specific WSD. The proposed technique can also be used for analysing the implications of including new potentially useful aspects to the WSD task in the biomedical domain, such as multilinguality [25]. In that work, information from multilingual corpora is added to the co-occurrence graphs used in the disambiguation process, for testing whether the use of smaller multilingual corpora is able to achieve similar results than those obtained through the use of big monolingual corpora.

Figure 1 illustrates the complete system, which we have named "Bio-Graph": In part **a)**, we can observe the creation of the knowledge base, which requires a preliminary annotation step. In

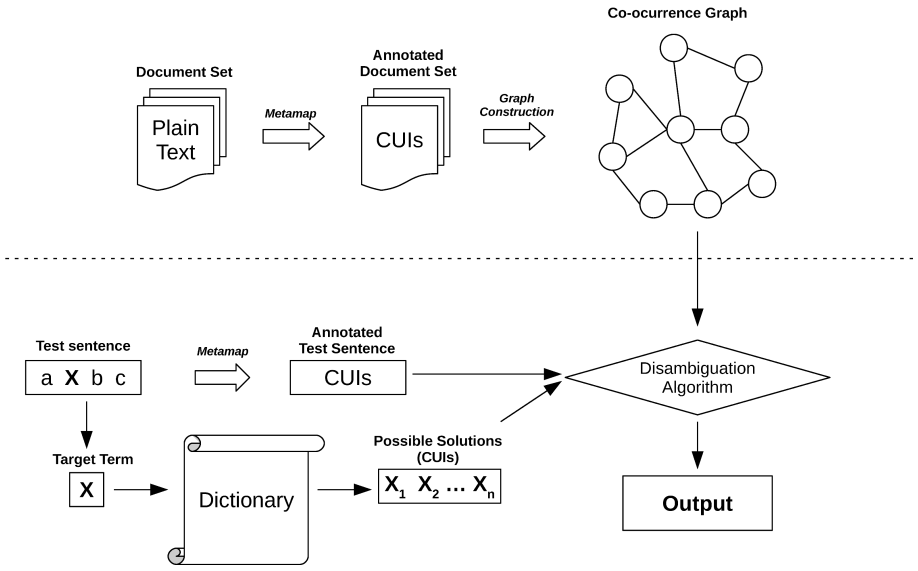

Figure 1: Construction of the co-occurrence graph (part **a**) and disambiguation of a test instance (part **b**).

this step, the text of each of the documents in the original set is transformed into medical concepts (UMLS CUIs). This new document set is then used for building the co-occurrence graph, through the statistical analysis that will be detailed later on. Part **b)** of the figure represents the disambiguation of a test instance. In this process, the ambiguous target term (represented by X in the figure) is located in the text, and its possible senses ($X_1, X_2, ..., X_n$) are extracted from a dictionary. Then, the text of the test instance is mapped onto CUIs. With this information (CUIs from context and possible senses) we can feed the co-occurrence graph and apply a disambiguation algorithm that will select, among those possible solutions, the most suitable sense of the ambiguous term in that context.

In this section, the annotation phase, as well as all the steps involved in the disambiguation, are detailed.

3.1. Annotation

The first step in the creation of the co-occurrence graph is to annotate the biomedical concepts that appear in the documents. These concepts will eventually become the nodes of the co-occurrence graph which forms the knowledge base used by our system. The annotation step consists in transforming the plain text that can be found in the medical documents, into CUIs that represent equivalent medical concepts. This step could be carried out by manual annotation, although in our case we perform it automatically, through the Metamap program [26], which allows us to split the text inside a document into phrases, and map each of those phrases onto a set of UMLS CUIs. This program offers the possibility of using a disambiguation server which helps the user to select a candidate for each phrase in the text. We make use of this server when annotating the documents that will be used for building the document graphs. Only unsupervised methods have been selected in the configuration of the disambiguation server, among those provided by the Metamap program, in order to maintain the unsupervised nature of the system throughout all

the process, while avoiding introducing too much noise to the co-occurrence graph. A baseline containing the results obtained by the disambiguation server considered in our experiments will be reported in subsequent sections. As we will see, the quality of this disambiguation is far from the results achieved in this work. We maintain the default values for the rest of the configuration parameters when running the Metamap program.

3.2. Graph Construction

The annotation step provides a set of documents, each of them containing a list of biomedical concepts represented by their UMLS CUIs. The next step is to determine the statistical significance of the co-occurrence of each possible pair of concepts inside this set of documents. For this purpose, we define a null model in which CUIs would be randomly and independently distributed among the documents of a corpus. We then compare the actual co-occurrences of each pair of CUIs against this null model (their probability of co-occurrence by pure chance) and select those that present a high statistical significance (low probability of being generated by the null model). More specifically, we calculate a p-value p for the co-occurrence of each pair of CUIs in our corpus. If p lies below a threshold next to 0, the co-occurrence is considered to be statistically significant, and hence those CUIs are considered to be related, and linked in the graph.

We consider two CUIs c_1 and c_2 appearing in n_1 and n_2 number of documents respectively (total number of documents is n). We calculate in how many ways those CUIs could co-occur in exactly k documents, by dividing the document collection in four different types of documents: k documents containing both c_1 and c_2, $n_1 - k$ documents containing only c_1, $n_2 - k$ containing only c_2, and $n - n_1 - n_2 + k$ containing neither c_1 nor c_2. The number of possible combinations is given by the multinomial coefficient:

$$\binom{N}{k, n_1 - k, n_2 - k} \tag{1}$$

The probability of those CUIs exactly co-occurring k times by pure chance is given by:

$$p(k) = \binom{N}{n_1}^{-1} \binom{N}{n_2}^{-1} \binom{N}{k, n_1 - k, n_2 - k} \tag{2}$$

if $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$ and zero otherwise.

To write equation (2) in a way that could be computationally more convenient, the notation $(a)_b \equiv a(a-1)\cdots(a-b+1)$ is introduced. For any $a \geq b$, and without loss of generality, we assume that $n_1 \geq n_2 \geq k$. Then,

$$\begin{aligned} p(k) &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2} (k)_k} \\ &= \frac{(n_1)_k (n_2)_k (N - n_1)_{n_2 - k}}{(N)_{n_2 - k} (N - n_2 + k)_k (k)_k}, \end{aligned} \tag{3}$$

where, in the second form, we used the identity $(a)_b = (a)_c (a-c)_{b-c}$, valid for any $a \geq b \geq c$. Finally, equation (3) can be rewritten as

$$p(k) = \prod_{j=0}^{n_2-k-1} \left(1 - \frac{n_1}{N-j}\right) \times \prod_{j=0}^{k-1} \frac{(n_1-j)(n_2-j)}{(N-n_2+k-j)(k-j)}. \quad (4)$$

The following p-value p for the co-occurrence of two CUIs can now be defined:

$$p = \sum_{k \geq r} p(k), \quad (5)$$

where r is the number of documents of our actual corpus in which we can find c_1 and c_2 together. As we stated before, if p lies below a determined threshold next to 0, the co-occurrence is statistically significant and a link between c_1 and c_2 is created in the graph. P-values of 0.01 and 0.05 are conventionally used when testing statistical significance. We have used a p-value of 0.01 for all the experiments described here. We also carried out additional analysis of the behaviour of our approach when more restrictive p-values are used (see Section 5.4). The weight of the link between two nodes i and j can be quantified in a practical way by defining it as $w_{ij} = \log(p_0/p_{ij})$, where p_0 is the selected threshold for the co-occurrence graph and p_{ij} is the p-value calculated using equation 5 and defining r as the actual number of co-occurrences between nodes i and j. Hence, the weight of the link will be proportional to the order-of-magnitude difference between p and p_0.

It is important to notice that the approach described here has the advantage that it does not assume that word frequencies are normally distributed, unlike some alternative measures of lexical co-occurrence [27]. For example, a chi-squared method would assume data to follow a gaussian distribution, which is not valid for many cases, especially when the number of co-occurrences is small. Our data only approximate gaussian for very large values, so chi-squared would not be recommended in this case. Hence we directly calculate how our actual data deviate from the null model proposed.

3.3. Disambiguation

Once that we have built our co-occurrence graph, we need to define a disambiguation algorithm. This algorithm will allow us to determine the most suitable sense (CUI) of an ambiguous concept (acronym or term) given its context, among all the possible senses provided by a dictionary. In other general WSD tasks, the selection or construction of this dictionary is a key point for assuring the good performance of a system [28]. In this particular task, the dictionary that contains the possible senses of every target word is publicly available.

The disambiguation algorithm that we have selected for performing this last step is the Personalized PageRank algorithm, initially introduced in [29]. This algorithm is based on the PageRank algorithm [30] which has been successfully applied to WSD tasks [31]. The PageRank algorithm is used over a graph for ranking the importance of each of its nodes. It is based on the relative structural importance of each node of the graph, represented by its incoming and outgoing edges.

The algorithm models, for each node, the probability of a random surfer over the graph ending on it. PageRank values for the whole graph can be calculated through the following formula:

$$P = cMP + (1 - c)v, \tag{6}$$

where P is the vector that contains the PageRank values for each node, c is a constant called "damping factor" usually set to 0.85, M is the matrix containing the values of the out-degrees of the nodes and v is a $N \times 1$ stochastic vector, being N the number of nodes in the graph. In this work, we will maintain the default value of the damping factor, this is, $c = 0.85$. Hence, the first element of the formula represents the movement of the random surfer between connected nodes, and the second one its probability of teleporting to any node without following the edges of the graph. By means of v, the probability of randomly jumping into a node of the graph can be distributed among the nodes of the graph in different ways. The Personalized PageRank approach makes use of this vector v for assigning higher probabilities to specific nodes of the graph. These probabilities will then spread along the graph, resulting in higher PageRank values for those nodes more influenced by the initial nodes highlighted in v.

In this case, the nodes that will be powered up in vector v are those that represent CUIs that appear in the context of the target concept we want to disambiguate. Hence, before performing the disambiguation step, we need to convert the plain text of each test instance onto the set of CUIs that represent all the medical concepts that can be found in the text, also using the Metamap program. When a term in the text is ambiguous, Metamap assigns all the possible CUIs that may correspond to it. When it comes to a target concept, this set of possible CUIs becomes the ambiguity that our system is trying to solve, since no disambiguation is selected in the Metamap program in this step. The rest of the configuration parameters in Metamap are set to their default values.

Once that we have all the CUIs that belong to the context of the target concept, we build v as a $N \times 1$ vector whose values will be $v_i = \frac{1}{C}$ if node i represents a CUI of the context, and 0 otherwise, being C the total number of CUIs found in the context of the target concept. After performing the Personalized PageRank algorithm, we will select the node with highest rank, among those representing possible senses of the target concept.

3.4. Example of Disambiguation

In this section an example of successful disambiguation illustrates the behavior of the Personalized PageRank (PPR) disambiguation on our co-occurrence graph, and compares it with the result obtained by running PPR over a graph directly built from the UMLS database. In this UMLS graph, two nodes are linked together if a relation between them can be found in the UMLS database. Figure 2 shows this example divided in two parts: the top part of the figure presents a test instance which contains the target word "culture", to be disambiguated. A look-up to the dictionary tells us that the two different senses (CUIs) of "culture" between which our system should discriminate are "C0430400", referred to a microbial culture (laboratory process), and "C0010453", referred to a culture from an anthropological point of view. Then, we obtain all the CUIs that represent concepts from the context of the test instance by applying Metamap to the text.

The second part of the figure (bottom part) illustrates the differences of applying the disambiguation process using our co-occurrence graph, or the UMLS graph. In our co-occurrence graph

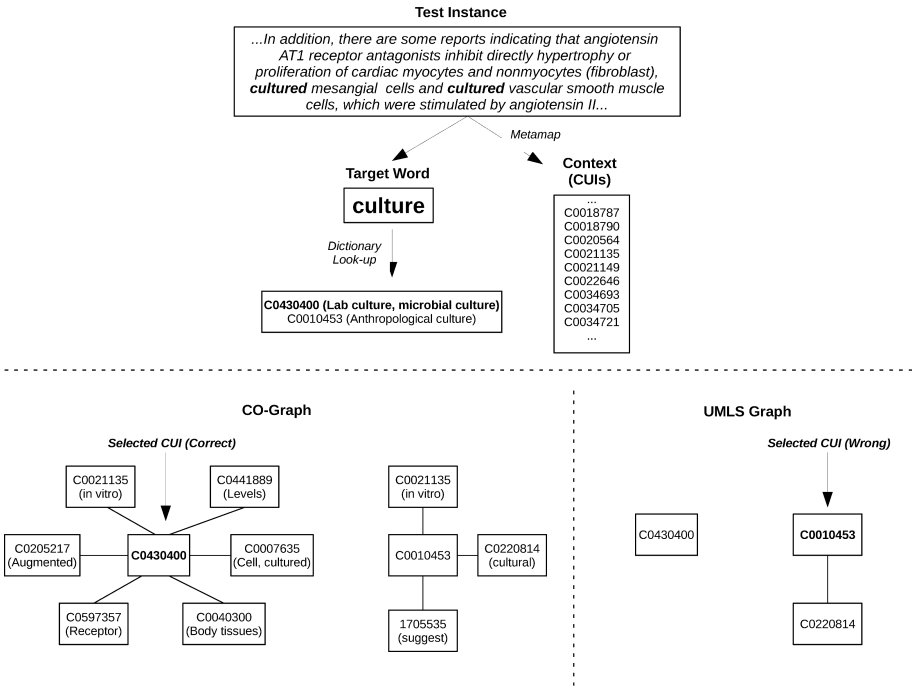

Figure 2: Example of disambiguation. Extraction of the target and context CUIs (top part) and comparison between the disambiguation algorithm over the co-occurrence graph and the UMLS graph (bottom part).

the correct sense of "culture" ("C0430400") is much more related to context CUIs than the other sense ("C0010453"). Hence, the disambiguation algorithm selects this more connected sense to be the most appropriate for this test instance. However, when using the UMLS graph we can observe that both senses are poorly connected to the CUIs in the context (which will result in a higher randomness when selecting a sense). In fact, the wrong sense is connected to one CUI in the context, while the correct sense is not connected to any of the CUIs in the context. Because of that, the disambiguation algorithm mistakenly selects the CUI "C0010453" to be the most appropriate for this test instance.

4. Datasets

This section describes the datasets used to evaluate our system.

4.1. Acronym Corpus

The Acronym corpus [32] contains 55,655 abstracts downloaded from Medline. Each of these abstracts contains an ambiguous acronym from a set of 21 originally developed in [33] and widely used in previous research. These acronyms each consist of at least 3 letters and are associated with between 2 and 5 extended forms (which are considered as senses). The dictionary for the target concepts is then created using the CUIs that correspond to their possible extended forms. The corpus is split into three different test datasets, containing 100 instances, 200 instances and 300 instances per ambiguous acronym, respectively. We will refer to those datasets as "A100",

"A200" and "A300". However, not all the 21 acronyms are present in every dataset, since some of them were removed from the test datasets due to an insufficient number of instances in the main corpus. Also, some acronyms such as "ACE", "ASP" and "CSF" were also removed from the initial datasets, in order to reduce their imbalance, since most of their test instances belonged to the same extended form. As a result, the A100, A200 and A300 datasets contain 18, 16 and 14 different ambiguous acronyms respectively. The final dataset obtained after this pre-processing is the same used by other state-of-the-art techniques to which we compare our system.

Data acquisition: Since the corpus was initially created for a supervised system, all the abstracts are annotated with the extended form that corresponds to the acronym found in the text. In this work, we present an unsupervised system that does not need these annotations, however, we need to acquire data to build our co-occurrence graph. This data will be represented by the abstracts from the original corpus that are not included in any of the three test datasets. Hence, our co-occurrence graph will be created from a set of 50,143 abstracts, which will be previously mapped onto CUIs from the UMLS database, as explained in Section 3.1.

4.2. NLM Corpus

The second corpus we will use to evaluate the performance of our system is the NLM-WSD corpus [34]. In contrast to the Acronym corpus, this corpus is composed of general ambiguous terms. It contains 50 terms with 100 instances per term. These instances are also abstracts downloaded from Medline, and manually annotated with the CUI that represents the correct sense for the target term in each instance. However, during the creation of the corpus, annotators could select to mark as "None" those instances for which none of the possible senses applied. We have removed those instances, so the final test dataset, which will be referred to as "NLM", contains 3,983 instances and 49 terms (since all the instances were marked as "None" for the term "association"). As with the Acronym corpus, this the same pre-processing is applied to the state-of-the-art techniques against which our system is compared.

Data acquisition: In this case, given that the NLM-WSD corpus is a test dataset itself, we do not have a set of documents to build the co-occurrence graph. Accordingly, we downloaded our own set of abstracts from Medline, using the Entrez interface [35]. We performed a search for each ambiguous term of the test dataset, restricting the results to 1,000 abstracts per term. In order to avoid downloading abstracts that could appear in the test dataset, we have only downloaded abstracts from year 2014. For maintaining the unsupervised nature of our technique, we do not specify in any way the sense of the ambiguous term for performing the search, so in the downloaded abstracts any possible sense of the target term can be found. The total number of abstracts in this set is 35,282. Although we downloaded 1,000 possible abstracts for each of the 50 ambiguous terms in the dataset, there are abstracts containing more than one term, and hence the reduction of the number of documents.

4.3. Dataset Properties

Table 1 resumes the characteristics of the datasets used for evaluation.

We can observe that for datasets "A200" and "A300" there is one abstract missing (given the number of ambiguous terms, and instances per term, they should have 3,200 and 4,200 instances respectively). This missing abstract was no longer available for download from Medline. The

	A100	A200	A300	NLM
Instances	1,800	3,199	4,199	3,983
Amb. terms	18	16	14	49
Min/Max # senses	2 / 4	2 / 4	2 / 4	2 / 5
Avg # senses	2.61	2.5	2.57	2.24

Table 1: Statistics for the different test datasets: number of instances, number of ambiguous terms (or acronyms), minimum and maximum number of senses for a term and average number of senses per term.

average number of possible senses is higher in the Acronym corpus than in the NLM corpus, although the total number of ambiguous terms is quite higher in this last corpus.

5. Evaluation

This section presents the results obtained by the approach described here and compares them with other state-of-the-art systems. An exhaustive analysis of the parameters of the system is also performed, in order to study how the results vary depending on their values.

5.1. System Results and Comparison

As we stated in previous sections, a co-occurrence graph was built for each of the evaluation corpus: the Acronym corpus (whose graph was used for evaluating the three test datasets, "A100", "A200" and "A300") and the NLM-WSD corpus. The performance metric used to evaluate system performance in all experiments is accuracy: number of correctly disambiguated instances divided by the total number of instances in the test dataset, expressed in %. Table 2 shows the accuracy achieved by our system in each of the test datasets. In order to analyse the impact of the selected co-occurrence graph when evaluating the system, we have also included the results obtained by cross-testing our graphs, this is, using the graph created with abstracts from the Acronym corpus for evaluating the "NLM" dataset, and vice versa. Finally, a joint graph was created combining the 50,143 abstracts of the "non-test" Acronym corpus and the 35,282 abstracts of the acquired "NLM-WSD related" corpus. The results of applying this joint graph to all the test datasets are also shown in the table.

	Datasets			
	A100	**A200**	**A300**	**NLM**
Acronym Graph	82.11	79.87	**82.64**	74.24
NLM Graph	61.83	59.59	58.83	75.45
Joint Graph	**82.78**	**80.06**	82.57	**78.36**

Table 2: Results (accuracy in %) for the co-occurrence graph-based system, for each of the graphs (Acronym corpus, NLM-related acquired corpus and joint graph), in each of the different test datasets. Bold highlights the best result obtained for each of the test datasets.

Results show that the graph created with abstracts from the Acronym corpus produces similar results on the three acronym test datasets. Regarding the cross-testing experiment, the results obtained using the Acronym-based graph over the NLM dataset are similar to those obtained by

using the NLM-based graph over the NLM dataset. However, the NLM-based does not perform as well in the cross-testing scenario, i.e. when applied to the Acronym datasets. This may be due to a greater specificity of the Acronym corpus, in which the different CUIs among which the disambiguation algorithm has to choose (representing extended forms of the acronyms), correspond to more specific concepts. On the other hand, terms in the NLM-WSD corpus are much more general. Hence, it is possible that some of the target CUIs of the Acronym corpus do not even appear in the graph created from NLM-related abstracts. Also, it is likely that any graph created from a large enough set of abstracts (such as the one created with acronym-based abstracts) contains enough information about CUIs representing the general concepts of the "NLM" dataset to perform a good disambiguation. Finally, we can observe that results obtained with the joint graph improve those obtained with simpler graphs for all but one of the datasets. This suggests that the combined information that can be found inside the joint graph is useful to better represent the connections between concepts and hence help to improve the overall disambiguation. We have conducted some additional experiments comparing the accuracy obtained using either the NLM-based Graph or the Joint Graph, both built with the same number of documents, and the achieved results confirm this intuition: 75.42% of accuracy of the Joint Graph against 69.12% of the NLM-based Graph for 10,000 documents, 77.45% against 71.48% for 20,000 documents, and 77.78% against 72.93% for 30,000 documents.

5.2. Comparison with Previous Approaches

Table 3 shows a comparison between the results obtained with our co-occurrence graph-based system ("Bio-Graph" in the table) as well as other knowledge-based and unsupervised systems that present results for the same datasets. The "NLM" dataset is more commonly used for evaluation than the Acronym datasets in the literature.

The first two rows of the table show results obtained using two different baselines: in the first row, we have the "Most Frequent Sense" (MFS) approach, which can be considered as a supervised baseline, and represents the accuracy achieved by a system that classifies every instance as belonging to the most common CUI for its ambiguous term. As we can observe, the MFS value for the NLM dataset is high demonstrating that it is imbalanced (i.e. for many of the ambiguous terms most of the instances belong to the same CUI). Also, we show results obtained by running the Metamap program against the test dataset, and making use of the disambiguation server under the same conditions we used for annotating the documents when building the co-occurrence graph, as explained in Section 3.1. As we can observe, the results for the NLM dataset are quite low in comparison with the accuracy achieved by our system. Since the Metamap program does not offer any disambiguation for acronyms, this second baseline does not offer results for the A100, A200 and A300 datasets.

Results from our system are compared against different WSD systems, mentioned in Section 2: The **PPR+UMLS** system [21] uses a graph-based similar approach, which makes use of a fixed graph built from the UMLS database, as described in the example shown in Figure 2. Although in the original work it is only applied to the "NLM" dataset, we have also reproduced this technique for testing the Acronym datasets, in order to obtain a better comparison. The **AEC** (Automatic Extracted Corpus) system [36] is a semi-supervised approach that automatically downloads and annotates abstracts for training a machine learning system. The **JDI** (Journal Descriptor Indexing)

	Datasets			
	A100	**A200**	**A300**	**NLM**
MFS	69.00	69.10	68.70	84.71
Metamap	—	—	—	49.13
PPR+UMLS	56.33	56.99	58.02	68.10
AEC	—	—	—	68.36
JDI	—	—	—	74.75*
MRD	—	—	—	63.89
2MRD	**88.00**	**90.00**	**89.00**	55.00
Bio-Graph	82.78	80.06	82.57	**78.36**

Table 3: Comparative of results (accuracy in %) for state-of-the-art systems (see text) and the system reported in this work (Bio-Graph), for each of the different test datasets. Bold highlights the best unsupervised results obtained for each of the test datasets.

method [37] makes use of semantic type vectors that represent each possible sense of an ambiguous term and computes their distance to a vector representing the test instance. Although it obtains good results for the NLM corpus, it only takes into account those senses belonging to different semantic types, hence many instances of the NLM corpus were removed in this experiment. That is the reason why results obtained by this system are marked with an asterisk in the table. Finally, the **MRD** and **2MRD** techniques are applied in [38] and [39] over the NLM corpus, while results achieved by the 2MRD technique over the Acronym datasets are presented in [22].

As we can observe in the table, our system outperforms all the state-of-the-art knowledge-based and unsupervised methods when applied to the NLM dataset, and even semi-supervised ones. Regarding the improvements obtained by our method with respect to the one that uses relations from the whole UMLS graph (PPR+UMLS), which can be considered the most similar approach to ours, we consider that contextual information obtained from actual abstracts in the process of building the graph is able to better represent knowledge that may eventually lead to correctly disambiguate a term inside a different abstract. Relations from the UMLS graph can be useful, but they do not necessarily imply that two related terms are likely to co-occur in the same document. The second-order vector technique (2MRD) outperforms our system in the Acronym corpus. However, while this technique makes use of additional information from UMLS (extended definitions of the possible senses), the main contribution of our method is that our disambiguation phase is completely based on the co-occurrence graph created from the abstracts, so it does not need additional information from the UMLS database.

Table 4 shows the word by word analysis of results for the NLM-WSD test dataset, for all the analysed methods, except the 2MRD system, for which we have not found those detailed results.

The Bio-Graph system is able to overcome the other systems in 24 out of 49 cases. The JDI system offers the best result in 12 cases, AEC and MRD are able to achieve the best result in 8 cases, and finally the Personalized PageRank technique over the UMLS graph outperforms the rest of the systems for 5 particular words. These results prove the strength of our technique across most of the words in the test dataset, which eventually leads to the best overall accuracy achieved by Bio-Graph.

Tables 5, 6 and 7 present a detailed description of the results obtained by the analysed systems

	NLM-WSD					
Word	**PPR+UMLS**	**AEC**	**JDI**	**MRD**	**Bio-Graph**	**Baseline (MFS)**
adjustment	0.3550	0.6237	**0.6923**	0.2308	0.6882	0.6667
blood pressure	0.4800	0.3700	0.2020	0.4343	**0.5000**	0.5300
cold	0.2840	0.3895	N/A	0.6044	**0.7579**	0.9053
condition	0.4890	0.7065	0.8370	0.3370	**0.9783**	0.9783
culture	0.7700	0.6000	**0.9700**	0.8200	0.9500	0.8900
degree	0.9380	0.8923	0.7077	0.4923	**0.9692**	0.9692
depression	0.9410	0.9529	0.9176	0.9941	**1.0000**	1.0000
determination	0.9490	0.1392	1.0000	**0.9936**	0.9494	1.0000
discharge	0.6930	0.7067	0.5556	**0.9861**	0.8400	0.9867
energy	0.2760	0.4000	0.7732	0.4536	**0.8200**	0.9900
evaluation	0.5000	0.5000	**0.5800**	**0.5800**	0.5000	0.5000
extraction	0.2760	0.7471	**0.9535**	0.2907	0.8621	0.9430
failure	0.7240	0.8621	**1.0000**	0.5862	0.1379	0.8621
fat	0.9590	0.8356	0.9296	**0.9718**	0.0274	0.9726
fit	0.1110	0.8889	**1.0000**	0.8387	**1.0000**	1.0000
fluid	**0.9200**	0.4800	0.3608	0.6082	0.8600	1.0000
frequency	0.9890	0.6064	0.1809	0.9362	**1.0000**	1.0000
ganglion	0.6400	0.8600	0.9130	**0.9565**	0.9300	0.9300
glucose	0.9000	0.7800	0.7347	0.2755	**0.9100**	0.9100
growth	0.3700	0.3700	0.6500	**0.6700**	0.6200	0.6300
immunosuppression	0.6200	0.5700	0.7083	0.4896	**0.7300**	0.5800
implantation	0.8470	**0.9490**	0.9053	0.8316	0.8673	0.8265
inhibition	0.2220	0.8384	**0.9899**	0.9697	**0.9899**	0.9899
japanese	0.6460	0.6329	0.8947	0.9211	**0.9241**	0.9367
lead	**0.9310**	0.8276	0.1724	0.3793	**0.9310**	0.9310
man	0.4460	**0.6522**	N/A	0.3187	0.6413	0.6304
mole	0.2740	0.4405	0.9398	0.8916	**0.9881**	0.9881
mosaic	0.6600	**0.8144**	0.7273	0.5795	0.4639	0.5360
nutrition	0.3260	0.3708	**0.4719**	0.3933	0.2697	0.5056
pathology	0.2830	0.6061	**0.8182**	0.3939	0.1717	0.8586
pressure	0.9790	0.5208	0.8172	**0.9836**	0.9688	1.0000
radiation	0.5310	0.7449	**0.7917**	0.6979	0.6224	0.6122
reduction	0.5450	**0.9091**	0.8182	0.8182	0.7273	0.8182
repair	0.7650	0.8529	0.8358	0.8358	**0.8971**	0.7647
resistance	0.6670	**1.0000**	**1.0000**	0.3333	0.0000	1.0000
scale	0.8460	0.7231	0.0615	0.0615	**0.9846**	1.0000
secretion	**0.9900**	0.4600	0.9798	0.3535	**0.9900**	0.9900
sensitivity	0.2750	0.7255	0.2745	0.8431	**0.9608**	0.9608
sex	0.8500	0.6000	N/A	0.5455	**0.8900**	0.8000
single	0.8200	0.8900	0.9300	0.0400	**0.9600**	0.9900
strains	0.9680	0.9570	1.0000	0.9780	**0.9785**	0.9892
support	0.8000	**1.0000**	0.9000	0.3000	0.2000	0.8000
surgery	**0.9700**	0.1900	0.8990	0.9394	0.7800	0.9800
transient	**0.9900**	0.9100	0.9600	**0.9900**	**0.9900**	0.9900
transport	0.6910	**1.0000**	**1.0000**	0.9780	0.9894	0.9894
ultrasound	0.8300	0.7400	0.7813	0.6667	**0.8400**	0.8400
variation	0.7500	0.6900	0.3500	0.7600	**0.8100**	0.8000
weight	0.5660	**0.6604**	N/A	0.4717	0.3208	0.5472
white	0.6330	0.5111	**0.6517**	0.4831	0.6444	0.5444
Accuracy all	0.6589	0.6836	0.7475*	0.6389	**0.7836**	0.8471

Table 4: Word by word comparative of results (accuracy in %) obtained by the analysed systems over the NLM-WSD test dataset, as well as by the Most Frequent Sense baseline (last column). Bold highlights the accuracy achieved by the best system for each of the words in the dataset, without considering the baseline. Last row shows the overall accuracy, with the best system also highlighted with bold typeface.

for each of the three Acronym test datasets (A100, A200 and A300), respectively. The accuracy achieved by the systems for each of the acronyms in the datasets is shown, as well as the overall accuracy obtained by each system.

	A100			
Acronym	**2MRD**	**PPR+UMLS**	**Bio-Graph**	**Baseline (MFS)**
ANA	0.8400	**0.8500**	0.7800	0.5800
APC	0.8800	0.7200	**0.9800**	0.3940
BPD	0.9600	0.3000	**0.9700**	0.4670
BSA	**0.9500**	0.8800	0.9400	0.8640
CAT	0.8800	0.5600	**0.9500**	0.5520
CML	0.8100	0.7500	**0.9200**	0.9170
CMV	**0.9800**	0.9700	**0.9800**	0.9670
DIP	**0.9800**	0.7900	0.9600	0.7510
EMG	**0.8800**	0.6900	0.1200	0.8840
FDP	0.6500	0.2300	**0.9500**	0.7850
LAM	0.8600	0.4800	**0.9600**	0.4830
MAC	**0.9400**	0.1500	0.6400	0.6430
MCP	**0.7300**	0.4100	0.6000	0.5020
PCA	0.7800	0.7200	**0.9700**	0.6890
PCP	0.9700	0.4200	**0.9900**	0.5780
PEG	0.8900	0.1600	**1.0000**	0.9410
PVC	**0.9500**	0.2300	0.2300	0.7820
RSV	**0.9700**	0.8300	0.9600	0.7670
Accuracy all	**0.8800**	0.5633	0.8278	0.6900

Table 5: Word by word comparative of results (accuracy in %) obtained by the analysed systems over the A100 test dataset. Bold highlights the accuracy achieved by the best system for each of the words in the dataset. Last row shows the overall accuracy, with the best system also highlighted with bold typeface.

As we can observe, the behaviour of all the systems is consistent across the three test datasets (A100, A200 and A300). Our system is able to obtain the best accuracy for most of the acronyms in the dataset. More specifically, Bio-Graph obtains the best result for 10 out of 18 acronyms for the A100 dataset, 8 out of 16 for the A200 dataset, and 8 out of 14 for the A300 dataset. The 2MRD system is only able to obtain the best accuracy for a similar number of cases in the A200 dataset. However, although our system presents high results for many particular cases, achieving good accuracy values for almost all the considered acronyms, it also presents very low accuracy for some cases (particularly, "EMG" and "PVC"), probably due to the nature of the corpus used for building the co-occurrence graph, which may suffer from lack of valuable information regarding those acronyms. This fact causes a lower overall accuracy when compared to the 2MRD system.

	A200			
Acronym	**2MRD**	**PPR+UMLS**	**Bio-Graph**	**Baseline (MFS)**
ANA	N/A	N/A	N/A	N/A
APC	0.8700	0.7500	**0.9650**	0.3940
BPD	0.9500	0.2650	**0.9800**	0.4670
BSA	**0.9300**	0.8800	0.9100	0.8640
CAT	0.8700	0.5600	**0.9500**	0.5520
CML	0.8400	0.7650	**0.9300**	0.9170
CMV	0.9800	0.9700	**0.9850**	0.9670
DIP	**0.9800**	0.7950	0.9600	0.7510
EMG	**0.8900**	0.7050	0.1150	0.8840
FDP	N/A	N/A	N/A	N/A
LAM	0.8700	0.4850	**0.9650**	0.4830
MAC	**0.9500**	0.1450	0.6550	0.6430
MCP	**0.6700**	0.4100	0.6150	0.5020
PCA	0.7900	0.7286	**0.9749**	0.6890
PCP	**0.9600**	0.4200	0.5800	0.5780
PEG	0.8900	0.1600	**1.0000**	0.9410
PVC	**0.9500**	0.2500	0.2500	0.7820
RSV	**0.9800**	0.8300	0.9750	0.7670
Accuracy all	**0.9000**	0.5699	0.8006	0.6910

Table 6: Word by word comparative of results (accuracy in %) obtained by the analysed systems over the A200 test dataset. Bold highlights the accuracy achieved by the best system for each of the words in the dataset. Last row shows the overall accuracy, with the best system also highlighted with bold typeface.

	A300			
Acronym	**2MRD**	**PPR+UMLS**	**Bio-Graph**	**Baseline (MFS)**
ANA	N/A	N/A	N/A	N/A
APC	0.8700	0.7633	**0.9600**	0.3940
BPD	0.9500	0.2600	**0.9767**	0.4670
BSA	**0.9200**	0.8700	0.9100	0.8640
CAT	0.8700	0.5633	**0.9367**	0.5520
CML	0.8300	0.7833	**0.9300**	0.9170
CMV	0.9800	0.9733	**0.9900**	0.9670
DIP	N/A	N/A	N/A	N/A
EMG	**0.8800**	0.7067	0.1167	0.8840
FDP	N/A	N/A	N/A	N/A
LAM	0.8800	0.4867	**0.9533**	0.4830
MAC	**0.9500**	0.1500	0.6433	0.6430
MCP	**0.6800**	0.4033	0.6267	0.5020
PCA	0.7900	0.7391	**0.9766**	0.6890
PCP	**0.9600**	0.4233	0.5767	0.5780
PEG	0.8800	0.1733	**0.9967**	0.9410
PVC	N/A	N/A	N/A	N/A
RSV	**0.9800**	0.8267	0.9667	0.7670
Accuracy all	**0.8900**	0.5802	0.8257	0.6870

Table 7: Word by word comparative of results (accuracy in %) obtained by the analysed systems over the A300 test dataset. Bold highlights the accuracy achieved by the best system for each of the words in the dataset. Last row shows the overall accuracy, with the best system also highlighted with bold typeface.

5.3. The effect of sense frequency: performance on the MSH-WSD dataset

The MSH-WSD dataset [39] is a test dataset also widely used in biomedical domain. It consists of 203 ambiguous entities (106 ambiguous abbreviations, 88 ambiguous terms and 9 combinations

of both). Apart from the inclusion of acronyms, the main difference with the NLM-WSD is the fact that its instances are very balanced. For each possible sense of each ambiguous term of abbreviation, the dataset contains approximately the same number of instances (around 100). In order to perform a more exhaustive evaluation of our method, we have also performed disambiguation experiments over the MSH-WSD dataset.

Similarly to the NLM-WSD dataset, all the documents in the MSH-WSD dataset have been annotated with the correct sense of the ambiguous term that they contain. Hence, the steps we have followed for acquiring data for building the co-occurrence graph have been the same as with the NLM-WSD dataset. This is, we have downloaded a number of abstracts for each ambiguous term, so each abstract can refer to any of the possible senses of the term (including those not considered in the MSH-WSD dataset). The total number of abstracts in this MSH-related corpus is 57,802.

Table 8 shows the accuracy of the already described unsupervised systems when applied to the MSH-WSD dataset, in comparison with the Bio-Graph system.

	MSH-WSD
MFS	54.50
AEC	**84.48**
JDI	65.51
MRD	81.18
2MRD	78.37
Bio-Graph	71.52

Table 8: Comparative of results (accuracy in %) for state-of-the-art systems (see text) and the system reported in this work (Bio-Graph), for the MSH-WSD dataset. Bold highlights the best unsupervised results obtained for each of the test datasets.

The table shows that the performance of the Bio-Graph system when applied to the disambiguation of the MSH-WSD dataset is lower than for other datasets. Regarding these results, some further analysis have been conducted on the MSH-WSD dataset, in order to determine the reason of the lower performance of Bio-Graph. Considering that our method relies on the use of a background corpus for building the co-occurrence graph, enough information should be found in this background corpus regarding all the possible senses of an ambiguous term, for the co-occurrence graph to correctly disambiguate it. Also, the number of documents containing each of the possible senses should be similar (this is, the background corpus should be properly balanced), otherwise the co-occurrence graph will be likely to present a bias towards the selection of those senses with higher presence in the corpus.

Following this intuition, a word-by-word analysis of these statistics has been performed, in order to compare the frequency of each ambiguous term and possible sense both in the Medline database and in the automatically acquired MSH-related corpus. This analysis is shown in Table 9.

The first five rows correspond to terms or abbreviations in the MSH-WSD dataset for which our system performs significatively worse than the abovementioned 2MRD system, selected for this analysis due to its overall higher similarity to the Bio-Graph system. The second five rows correspond to terms or abbreviations for which our system performs significatively better than the

	Systems		Medline		MSH-related	
Term	**Bio-Graph**	**2MRD**	**Term frequency (min)**	**Balance ratio**	**Term frequency (min)**	**Balance ratio**
AA	0.5025	**0.9899**	34,503 (249)	0.1707	523 (4)	0.0077
CCD	0.2979	**0.9929**	5011 (43)	0.3945	13 (2)	0.1818
Cortex	0.5076	**0.9495**	174,069 (291)	0.0077	807 (4)	0.0050
FTC	0.5533	**0.9848**	1312 (157)	0.5358	135 (4)	0.0305
Pneumocystis	0.4975	**0.8586**	8557 (849)	0.2108	70 (0)	0
Lactation	**0.8782**	0.6919	21,886 (1083)	0.0849	437 (197)	0.8208
Nurse	**0.8081**	0.6616	68,619 (291)	0.0194	686 (240)	0.5381
POL	**0.9506**	0.7346	9667 (65)	0.0980	242 (109)	0.8195
SARS	**0.7374**	0.5808	5481 (802)	0.4456	603 (289)	0.9204
Tolerance	**0.8485**	0.6717	122,691 (9348)	0.9294	208 (84)	0.6774

Table 9: Excerpt from the word by word comparative of results regarding the MSH-WSD dataset. Second and third columns show the accuracy obtained by our system and by the 2MRD system, respectively. Fourth and fifth columns show frequency statistics in Medline (see text). Sixth and seventh columns show frequency statistics in the MSH-related corpus (see text).

2MRD system. We compare the statistics of the terms in the whole Medline database, as well as in the MSH-related corpus that we created for evaluating our system in this dataset. For both cases (Medline and MSH-related), we show the overall term frequency, including in parentheses the frequency of the sense with fewer appearances, as well as the "balance ratio", computed as the ratio between the sense with higher frequency and the term with lower frequency, among those selected for the evaluation dataset. This metric will come closer to 1 as the distribution of frequencies of the senses of the given term is more balanced.

As we can observe, in those cases in which our system performs worse than the 2MRD system, both the minimum frequency and the balance ratio in the MSH-related corpus are very low, especially when compared to the minimum frequency and the balance ratio of the Medline database. On the other hand, those terms with better disambiguation accuracy present a more balanced distribution of senses across the MSH-related corpus, and also a higher minimum and overall frequency of appearance. This fact indicates that, for this particular dataset, and despite of obtaining a relatively high number of documents for each ambiguous term when building the MSH-related corpus, we are not gathering enough samples of some specific senses for the co-occurrence graph to correctly disambiguate the test instances that correspond to those senses. Hence, the annotation step should be revisited in order to deal with this issue.

5.4. Parameter Analysis

In this section we explore the effect of varying the two parameters used by the approach described here. The joint graph (built with abstracts from both the Acronym and the NLM-related corpus) is used for the experiments described here.

The first parameter is the threshold for the p-value p (see Section 3.2). This threshold, denoted by p_0, establishes the highest accepted value for p in order to consider a co-occurrence to be statistically significant, and hence create a link in the graph between the two co-occurring CUIs. Figure 3 illustrates the behaviour of our system, in terms of accuracy for each test dataset, when we vary p_0, decreasing its value from $p_0 = 10^{-2}$ to $p_0 = 10^{-11}$. As previously stated (Section 3.2), we have chosen a maximum value of 0.01. Experiments in which greater thresholds were used showed that the resulting graphs are unmanageably large and that performance quickly decreases.

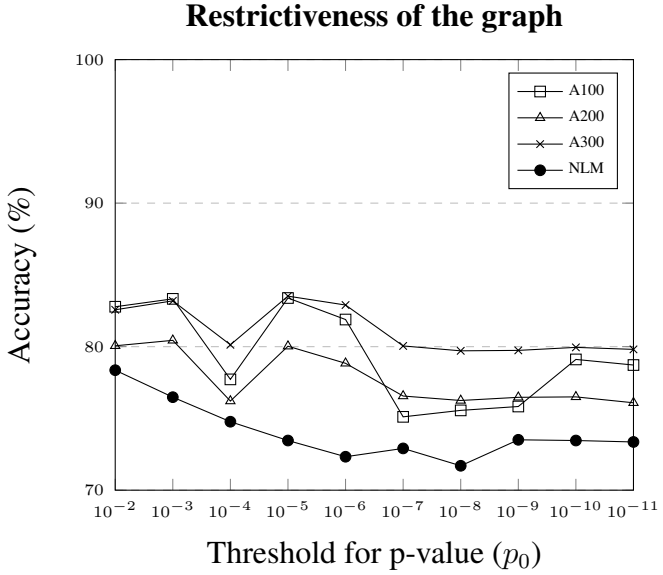

Figure 3: Evolution of the accuracy (%) as the specified threshold for the p-value decreases (the restrictiveness of the graph increases).

As we decrease the threshold, it is more difficult for a pair of CUIs to present a statistically significant p-value, and hence the graph becomes more restrictive, reducing the number of edges. The best results are obtained for the least restrictive graphs, while accuracy usually decreases as we decrease p_0. This is due to the removal of important edges representing relations between concepts, as we increase the restrictiveness of the graph.

Figure 4 represents the behaviour of the system depending on the number of abstracts used for building the co-occurrence graph. The complete set of abstracts used for building the joint graph was randomized, and gradually larger subsets of those abstracts were used to build the graphs. As we increase the number of abstracts, each subset contains all the abstracts of the previous one.

The overall accuracy increases with the number of abstracts used to build the graph, although performance for each method quickly reaches a plateau. Results rapidly converge to an accuracy of more than 80% in the A100, A200 and A300 datasets, and around 77% in the NLM dataset. Fast convergence of the algorithm is a useful feature when resources are limited.

6. Conclusions and Future Work

This paper describes the application of a technique based on co-occurrence graphs for performing WSD in the biomedical domain. The knowledge base on which the system relies is automatically created in an unsupervised way from a set of abstracts downloaded from the Medline database and automatically mapped onto medical concepts. Unlike other state-of-the-art techniques, external resources are not used for the disambiguation step. Evaluation on two widely used test datasets shows that the reported method obtains consistent results that outperform most of the knowledge-based systems addressing the same problem. Further experiments suggest that the convergence of the method is fast regarding the number of abstracts used for building the graph. In addition, better

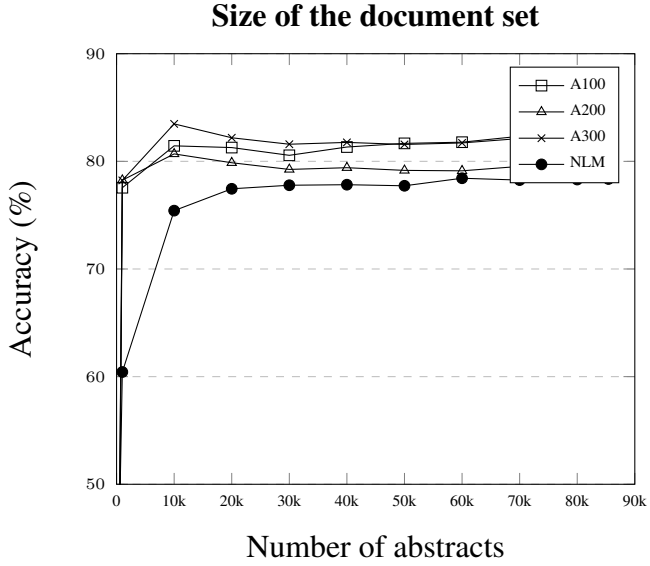

Figure 4: Evolution of the accuracy (%) as the number of abstracts used for building the co-occurrence graph increases.

results are obtained with less restrictive graphs, since they incorporate to the co-occurrence graph the most useful information about relations between concepts for performing the disambiguation.

Planned future work includes application of disambiguation algorithms that take into account weights of links of the graph. Some of these algorithms could be used to create communities (densely connected sub-graphs) of concepts that represent information about the possible senses of an ambiguous term in a more accurate way. Finally, combining our method with techniques derived from similar state-of-the-art systems may improve results further, especially for scenarios in which the performance of our system is somehow limited, such as the particular cases of the MSH-WSD dataset illustrated in Section 5.3.

7. Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the projects EXTRECM (TIN2013-46616-C2-2-R), PROSA-MED (TIN2016-77820-C3-2-R) and EXTRAE (IMIENS 2017), as well as by the Universidad Nacional de Educacion a Distancia (UNED) through the FPI-UNED 2013 grant.

References

[1] G. K. Savova, A. R. Coden, I. L. Sominsky, R. Johnson, P. V. Ogren, P. C. de Groen, C. G. Chute, Word sense disambiguation across two domains: Biomedical literature and clinical notes, Journal of Biomedical Informatics 41 (6) (2008) 1088 – 1100. doi:http://dx.doi.org/10.1016/j.jbi.2008.02.003.

[2] M. Stevenson, Y. Guo, Disambiguation in the biomedical domain: The role of ambiguity type, Journal of Biomedical Informatics 43 (6) (2010) 972 – 981. doi:http://dx.doi.org/10.1016/j.jbi.2010.08.009.

[3] A. K. Sehgal, P. Srinivasan, O. Bodenreider, Gene terms and english words: An ambiguous mix, in: Proc. of the ACM SIGIR Workshop on Search and Discovery for Bioinformatics, Sheffield, UK, Citeseer, 2004.

[4] J. Martinez-Romo, L. Araujo, J. Borge-Holthoefer, A. Arenas, J. A. Capitán, J. A. Cuesta, Disentangling categorical relationships through a graph of co-occurrences, Phys. Rev. E 84 (2011) 046108. doi:10.1103/PhysRevE.84.046108.
[5] M. J. Schuemie, J. A. Kors, B. Mons, Word sense disambiguation in the biomedical domain: an overview, Journal of Computational Biology 12 (5) (2005) 554–565.
[6] E. Agirre, P. G. Edmonds, Word sense disambiguation: Algorithms and applications, Vol. 33, Springer Science & Business Media, 2007.
[7] S. Moon, S. Pakhomov, G. B. Melton, Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations, in: AMIA Annual Symposium Proceedings, Vol. 2012, American Medical Informatics Association, 2012, p. 1310.
[8] Y. K. Lee, H. T. Ng, An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation, in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 41–48.
[9] I. Iacobacci, M. T. Pilehvar, R. Navigli, Embeddings for word sense disambiguation: An evaluation study, ACL, 2016.
[10] M. Joshi, S. Pakhomov, T. Pedersen, C. G. Chute, A comparative study of supervised learning as applied to acronym expansion in clinical reports, in: AMIA Annual Symposium Proceedings, Vol. 2006, American Medical Informatics Association, 2006, p. 399.
[11] H. Xu, M. Markatou, R. Dimova, H. Liu, C. Friedman, Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues, BMC bioinformatics 7 (1) (2006) 334.
[12] S. Moon, B.-T. Berster, H. Xu, T. Cohen, Word sense disambiguation of clinical abbreviations with hyperdimensional computing, in: AMIA Annual Symposium Proceedings, Vol. 2013, American Medical Informatics Association, 2013, p. 1007.
[13] Y. Wu, J. Xu, Y. Zhang, H. Xu, Clinical abbreviation disambiguation using neural word embeddings, ACL-IJCNLP 2015 (2015) 171.
[14] M. Stevenson, Y. Guo, Disambiguation of ambiguous biomedical terms using examples generated from the umls metathesaurus, Journal of biomedical informatics 43 (5) (2010) 762–773.
[15] H. Xu, P. D. Stetson, C. Friedman, Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations, in: AMIA Annual Symposium Proceedings, Vol. 2012, American Medical Informatics Association, 2012, p. 1004.
[16] C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books, 1998.
[17] R. S. Sinha, R. Mihalcea, Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity., in: ICSC, Vol. 7, 2007, pp. 363–369.
[18] R. Navigli, M. Lapata, An experimental study of graph connectivity for unsupervised word sensedisambiguation, IEEE transactions on pattern analysis and machine intelligence 32 (4) (2010) 678–692.
[19] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, G. O. Barnett, The unified medical language system, Journal of the American Medical Informatics Association 5 (1) (1998) 1–11.
[20] R. Chasin, A. Rumshisky, O. Uzuner, P. Szolovits, Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods, Journal of the American Medical Informatics Association 21 (5) (2014) 842–849.
[21] E. Agirre, A. Soroa, M. Stevenson, Graph-based word sense disambiguation of biomedical documents, Bioinformatics 26 (22) (2010) 2889–2896. doi:10.1093/bioinformatics/btq555.
[22] B. T. McInnes, T. Pedersen, Y. Liu, S. V. Pakhomov, G. B. Melton, Using second-order vectors in a knowledge-based method for acronym disambiguation, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, 2011, pp. 145–153.
[23] A. J. Yepes, R. Berlanga, Knowledge based word-concept model estimation and refinement for biomedical text mining, Journal of biomedical informatics 53 (2015) 300–307.
[24] A. Duque, L. Araujo, J. Martinez-Romo, Co-graph: A new graph-based technique for cross-lingual word sense disambiguation, Natural Language Engineering 21 (2015) 743–772. doi:10.1017/S1351324915000091.
[25] A. Duque, J. Martinez-Romo, L. Araujo, Can multilinguality improve biomedical word sense disambiguation?, Journal of biomedical informatics 64 (2016) 320–332.

[26] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: The metamap program, Proceedings of the American Medical Informatics Association (AMIA) (2001) 17–21.
[27] D. B. Hitchcock, Yates and contingency tables: 75 years later., Journal lectronique d'Histoire des Probabilits et de la Statistique [electronic only] 5 (2) (2009) 1–14; electronic only.
[28] A. Duque, J. Martinez-Romo, L. Araujo, Choosing the best dictionary for cross-lingual word sense disambiguation, Know.-Based Syst. 81 (C) (2015) 65–75. doi:10.1016/j.knosys.2015.02.007.
[29] T. H. Haveliwala, Topic-sensitive pagerank, in: Proceedings of the 11th International Conference on World Wide Web, WWW '02, ACM, New York, NY, USA, 2002, pp. 517–526. doi:10.1145/511446.511513.
[30] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, in: COMPUTER NETWORKS AND ISDN SYSTEMS, Elsevier Science Publishers B. V., 1998, pp. 107–117.
[31] E. Agirre, A. Soroa, Personalizing pagerank for word sense disambiguation, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 33–41.
[32] M. Stevenson, Y. Guo, A. Al Amri, R. Gaizauskas, Disambiguation of biomedical abbreviations, in: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, BioNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 71–79.
[33] H. Liu, Y. A. Lussier, C. Friedman, Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method, Journal of Biomedical Informatics 34 (4) (2001) 249 – 261. doi:http://dx.doi.org/10.1006/jbin.2001.1023.
[34] M. Weeber, J. G. Mork, A. R. Aronson, Developing a test collection for biomedical word sense disambiguation, in: Proceedings of the AMIA 2001 Symposium, 2001, pp. 746–750.
[35] E. Sayers, A general introduction to the e-utilities.
[36] A. Jimeno-Yepes, A. R. Aronson, Knowledge-based biomedical word sense disambiguation: comparison of approaches., BMC Bioinformatics 11 (2010) 569.
[37] S. M. Humphrey, W. J. Rogers, H. Kilicoglu, D. Demner-fushman, T. C. Rindflesch, Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment, J. Am. Soc. Inform. Sci. Tech 57 (2006) 96–113.
[38] B. T. McInnes, An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop, Association for Computational Linguistics, 2008, pp. 49–54.
[39] A. J. Jimeno-Yepes, B. T. McInnes, A. R. Aronson, Exploiting mesh indexing in medline to generate a data set for word sense disambiguation, BMC bioinformatics 12 (1) (2011) 223.