Journal of Biomedical Informatics 64 (2016) 320–332

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Can multilinguality improve Biomedical Word Sense Disambiguation?

Andres Duque, Juan Martinez-Romo, Lourdes Araujo

NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain

ARTICLE INFO

Article history:
Received 26 July 2016
Revised 24 October 2016
Accepted 31 October 2016
Available online 2 November 2016

Keywords:
Biomedical Word Sense Disambiguation
Multilinguality
Graph-based systems
Unified Medical Language System
Unsupervised systems
Parallel and comparable corpora

ABSTRACT

Ambiguity in the biomedical domain represents a major issue when performing Natural Language Processing tasks over the huge amount of available information in the field. For this reason, Word Sense Disambiguation is critical for achieving accurate systems able to tackle complex tasks such as information extraction, summarization or document classification. In this work we explore whether multilinguality can help to solve the problem of ambiguity, and the conditions required for a system to improve the results obtained by monolingual approaches. Also, we analyze the best ways to generate those useful multilingual resources, and study different languages and sources of knowledge. The proposed system, based on co-occurrence graphs containing biomedical concepts and textual information, is evaluated on a test dataset frequently used in biomedicine. We can conclude that multilingual resources are able to provide a clear improvement of more than 7% compared to monolingual approaches, for graphs built from a small number of documents. Also, empirical results show that automatically translated resources are a useful source of information for this particular task.

1. Introduction

There are multiple scenarios in the biomedical domain in which data scarceness is one of the major issues for building a system that successfully performs Natural Language Processing (NLP) tasks. This scenarios include studies developed in low-income or middle-income countries in which health research efforts and resources are unequally distributed [1], or works regarding low resource languages. Also, information related to specific tasks such as the study of rare diseases is scarce and difficult to summarize, as well as time-consuming for the few experts in the area [2]. Hence, we are talking about poorly documented problems, for which most of the available corpora in the literature will be small [3]. Widely explored in the NLP literature, multilinguality has been proven to be a really useful source of information when it comes to NLP tasks [4–6]. The use of multilingual data could palliate this lack of information in some fields of the biomedical domain. Hence, one of the initial hypothesis of this work considers that significant improvements can be achieved in NLP tasks in the biomedical domain by adding multilingual information to a knowledge-based system.

In particular, we focus on the Word Sense Disambiguation (WSD) task, in which the main objective is to solve the lexical ambiguity [7,8] of biomedical documents such as scientific papers or medical reports. Given a test sentence which contains an ambiguous term, a system should determine which of its possible senses is the most appropriate considering the context. For example, the word "surgery" may refer to the branch of medicine that applies operative procedures to treat diseases, or to one of those operative procedures. There exist many different types of lexical ambiguity in biomedical documents, which represents an additional challenge when performing WSD in this domain [9]: 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 (more than one thousand gene terms overlap with generic English meanings [10]). The use of biomedical concepts, in addition to plain text, when working with medical documents, can be seen as another challenge, since the process of transforming plain text into biomedical concepts is an additional step not considered when working with more general texts, that is, not belonging to a specific domain [11].

It is difficult to find works in the literature that apply multilinguality to the WSD task, probably due to the lack of bilingual corpora providing enough useful information for disambiguation, that is, a wide enough collection of documents containing ambiguous terms, and with a balanced number of occurrences for each possible sense of such terms. Yet, WSD is of paramount importance for many document processing tasks, such as summarization, text classification or information extraction. New possibilities of improvement related to WSD are thus highly relevant in the NLP field.

E-mail addresses: aduque@lsi.uned.es (A. Duque), juaner@lsi.uned.es (J. Martinez-Romo), lurdes@lsi.uned.es (L. Araujo)

http://dx.doi.org/10.1016/j.jbi.2016.10.020
1532-0464/

The main contribution of this paper is the application of multilingual techniques to an unsupervised graph-based approach for performing WSD in the biomedical domain, to analyze the improvements that can be achieved by this kind of data when evaluating this multilingual system on widely known datasets containing a range of ambiguities. We perform a thorough analysis of the conditions under which the proposed approach becomes a useful and powerful tool to solve the WSD problem. We also explore ways of dealing with the lack of available bilingual corpora, as well as different languages and their contributions to possible improvements.

The rest of the paper is organized as follows: Section 2 provides background on approaches regarding multilinguality for WSD, and WSD in the biomedical domain. Section 3 presents the system and algorithms used in this work, explaining in detail all the steps involved in the disambiguation process. Considerations about the test environment used for evaluation are presented in Section 4. The different experiments and the results obtained, as well as a detailed discussion for each of them are described in Sections 5–7 respectively. An example of behavior of the system and the disambiguation process is presented in Section 8. Finally, Section 9 contains the final conclusions and future lines of work.

2. Previous work

Multilinguality has been widely explored in NLP processes, and especially in WSD tasks, for which parallel corpora have been used as source of information, given their potentiality for disambiguation [12]. In that work, an evaluation framework and an approach for measuring the distance between senses were proposed. The translation of an ambiguous word from a language into another can offer very useful clues about its disambiguation in any of the studied languages. Other works have also exploited this kind of resources for automatically tagging senses using sense inventories for each of the languages in a corpus [13], or for implementing supervised models that make use of multilingual features extracted from the corpus [14]. Apart from parallel corpora, resources such as comparable corpora, which are not restricted to document-level or sentence-level alignments across languages [15], may also be good resources for this kind of tasks. The automatic generation of these multilingual resources has been also studied and compared with manually generated multilingual corpora [16]. WSD has been frequently addressed under a supervised point of view [17,18], originally through methods based on probabilistic models and their variants [19], as well as other machine learning algorithms [20], and in the last few years exploiting the development of word embeddings [21]. However, many unsupervised and knowledge-based techniques have been also developed in order to deal with the lack of annotated training information [22,23]. Particularly, graph-based techniques have been explored in this domain, although most of them make use of WordNet [24] as knowledge base for extracting the information used for disambiguation. For example, different semantic similarity measures are computed in [25] for generating a graph between the ambiguous target term and its surrounding words, depending on the importance of the relationships between senses of the target word and senses of the context words, given by those similarity measures. The disambiguation is performed by running different centrality algorithms over the resulting graph, hence highlighting the most suitable sense. Other similar works make use of the different lexical chains (sequences of semantically related words) that can be found in the text containing the ambiguous word, for building graphs which will eventually help in the disambiguation process, through different algorithms [26,27].

Regarding the biomedical domain, WSD is a key step to automatically access, retrieve and process the increasing amount of available unstructured textual information [11]. WSD processes should be implemented in almost every system attempting to perform more complex NLP tasks, such as summarization [28] or information extraction [29]. As in general WSD, it is commonly accepted that most of the systems performing biomedical WSD can be separated into two main groups: data-driven or algorithms that need labeled training data, and knowledge-based techniques [30]. Many works performing supervised WSD can be found in the literature, most of them making use of linguistic features that are usually employed for performing WSD in more general domains [9]. Features such as part-of-speech (POS) tags, unigrams and bigrams are used in [31] for training Naïve Bayes classifiers, decision trees and Support Vector Machines (SVMs) and their results are compared. SVM is also used in [32] for abbreviation disambiguation. Vector Symbolic Architectures (VSA) have been used in [33] for encoding vector representations for the ambiguous term and each of its senses. This representation can be reversed for new instances containing the ambiguous term in order to recover the appropriate sense for the context. More recent works have also applied state-of-the-art deep learning techniques such as neural word embeddings to acronym disambiguation [34]. In this work, different techniques are implemented for deriving word embeddings of ambiguous terms, and their performance is compared inside a system which uses a SVM algorithm, taking the word embeddings as inputs. Also semi-supervised works which introduce "pseudo-data" [35] or create automatically extracted and annotated training corpus for building Machine Learning (ML) systems [36] present successful results in the considered task.

In the present work we are going to explore knowledge-based approaches, based on initially untagged corpora and external resources such as the Unified Medical Language System (UMLS) Metathesaurus [37]. In this database, biomedical concepts are unequivocally represented through Unique Concept Identifiers (CUIs), and linked together depending on their relationships [38]. Knowledge-based methods usually take advantage of this representation and the additional information that it provides for performing the disambiguation [39,40]. Other knowledge-based systems will be defined and compared against the proposed system in subsequent sections. As far as we have been able to find in the literature, the use of multilingual information for performing WSD in the biomedical domain has not been explored, and hence the main motivation of this work is to analyze this field of study and offer results either encouraging or advising against the use of multilingual information.

3. System description

The multilingual technique described in this paper makes use of a co-occurrence graph as source of knowledge for performing WSD. This co-occurrence graph, whose theoretical background is described in detail in [41], is based on the hypothesis that documents inside a corpus are consistent, that is, 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 [42], with successful results, which suggests that a similar approach could also lead to competitive results in biomedical WSD. Also, the success of the approach when tested on different WSD problems allows us to rely on the ade-

quacy of using the proposed co-occurrence graph on this new problem.

In this particular work, we consider two types of concepts that may eventually become nodes of the co-occurrence graph: First of all, we have specific medical concepts that can be found in the UMLS database, and are identified through their CUIs. This identifier is the required output of a system that performs WSD in the biomedical domain, since it unequivocally represents a specific sense. Hence, this information is crucial to exactly determine which sense is the most appropriate for an ambiguous word given its context. However, the UMLS database is mainly restricted to the English language, and hence we need to define another type of concepts which represent the additional information given by other languages, and even by the English language. This second type of concepts are words in the documents, carefully annotated and filtered in order to eliminate all the non-informative words. In particular, we will focus on nouns and adjectives for considering informative words and avoiding introducing too much information into the knowledge base (the co-occurrence graph), which may lead to unmanageable graphs.

Fig. 1 illustrates the complete system: In part (**a**), we can observe the creation of the knowledge base, which requires a preliminary annotation step. For this step, we have documents written both in English and in any other language which will be used for enriching the knowledge base. The text of each of the documents in the original set is transformed into medical concepts (UMLS CUIs), and nouns and adjectives from English and other languages are extracted. 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. First, the test instance has to be translated into every language in the multilingual corpus. The ambiguous target term (represented by X in the figure) is located in the text, and its possible senses $(X_1, X_2, \ldots, X_n)$ are extracted from a dictionary. Then, the text of the original test instance, written in English, is mapped onto CUIs. Also nouns and adjectives are extracted from the English sentence and the translated ones. With this information (CUIs and textual information) 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 the next subsections, 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. As we stated before, we have two types of concepts: CUIs and words. In order to filter all the non-informative words out of the text documents, we need to lemmatize and tag those documents with Part-Of-Speech (POS) tags. This procedure is automatically performed by the TreeTagger tool [43], both for English and for the other languages in the multilingual corpus considered in this work. For generating the other kind of concepts that we want to include in our co-occurrence graph (UMLS CUIs), we need to transform the plain text written in English 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 [44], which allows us to split the text inside a document written in English 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. If it was not used, each time an ambiguous medical concept appeared in a document it would be replaced by all its possible senses (CUIs) and consequently the co-occurrences would not provide us useful information for disambiguation. However, as the configuration of the disambiguation server can be set when running the program, we have selected unsupervised methods for this initial disambiguation in all the experiments reported in this work. This way, we assure that the unsupervised nature of our system is maintained through all the process. 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

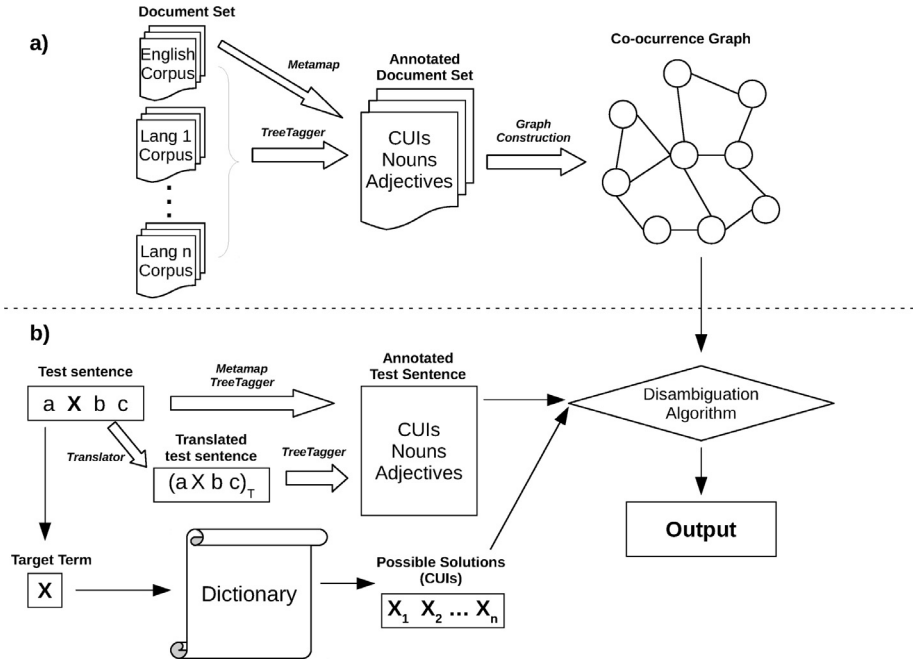

Fig. 1. Construction of the co-occurrence graph (part **a**) and disambiguation of a test instance (part **b**).

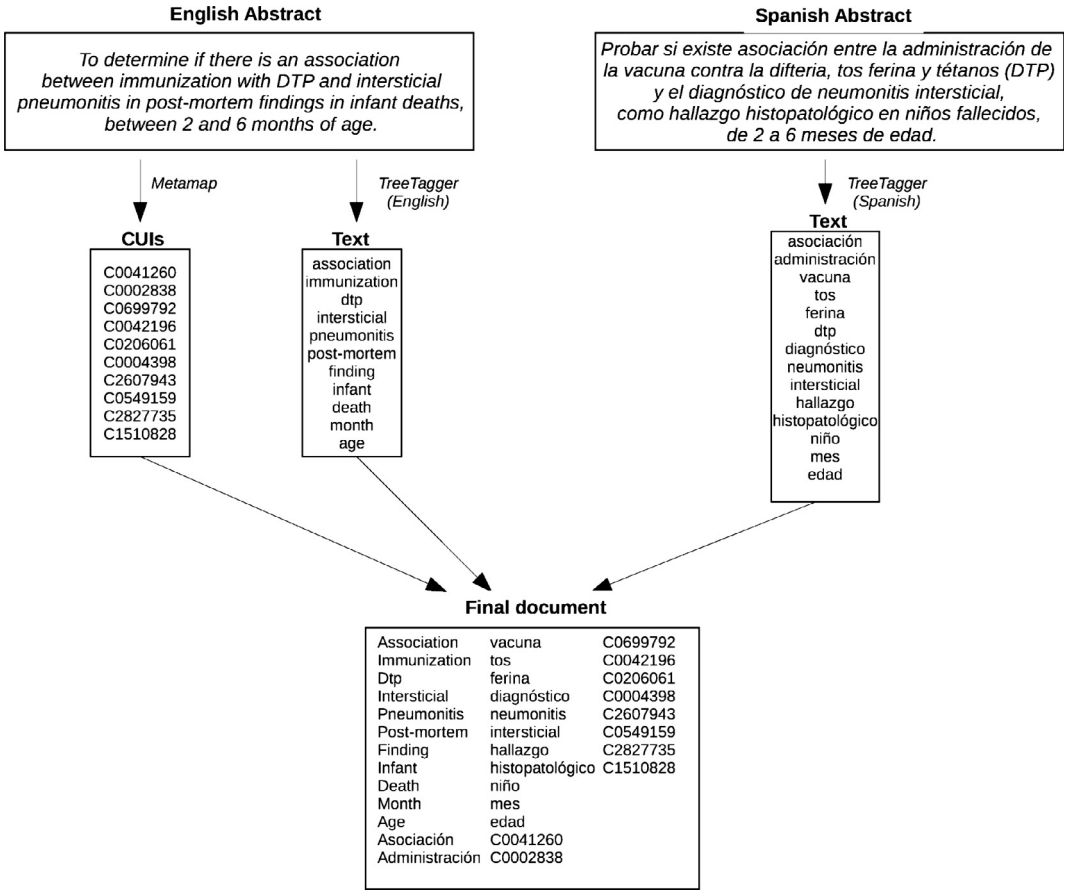

Fig. 2. Example of annotation of a test instance written in English and Spanish. CUIs from the Metamap-annotated English document, and nouns and adjectives from both languages are joined together into the final document, which contains concepts for populating the co-occurrence graph.

maintain the default values for the rest of the configuration parameters when running the Metamap program.

Fig. 2 shows an example of the annotation step, for the excerpt of an abstract written in English and Spanish. We observe the process of annotating the English text with the Metamap tool for extracting the CUIs. Also, both documents are annotated with TreeTagger for obtaining nouns and adjectives. The final document contains all the concepts that may eventually become nodes of the co-occurrence graph.

3.2. Graph construction

The annotation step provides us with a set of documents, each of them containing a list of biomedical concepts represented by their UMLS CUIs, as well as nouns and adjectives in both English and the additional language or languages we use for enriching our co-occurrence graph. The next step is to determine the statistical significance of the co-occurrence of each possible pair of concepts (either CUI or word) inside this set of documents. For this purpose, we define a null model in which concepts would be randomly and independently distributed among the documents of a corpus. We then compare the actual co-occurrences of each pair of concepts 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 concepts in our corpus. If p lies below a threshold next to 0, the co-occurrence is considered to be statistically significant, and hence those concepts are considered to be related, and linked in the graph.

We consider two concepts 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 concepts 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 concepts 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 $\max\{0, n_1 + n_2 - N\} \leqslant k \leqslant \min\{n_1, n_2\}$ and zero otherwise.

To write Eq. (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 $a \geqslant b$, and without loss of generality we assume that $n_1 \geqslant n_2 \geqslant k$. 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 $a \geqslant b \geqslant c$. Finally, Eq. (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 concepts can now be defined:

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

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 specific threshold p_0 next to 0, the co-occurrence is statistically significant and a link between c_1 and c_2 is created in the graph. The maximum p-value for statistical significance purposes usually found in the literature is $p_0 = 0.01$ or $p_0 = 0.05$ (99% or 95% of statistical trust). In this work, we have defined a maximum p-value (threshold) of $p_0 = 0.01$ for all the experiments described.

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 [45]. 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 correspond to a hypergeometric distribution (which only approximates 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 hypergeometric distribution (null model).

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 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 [46]. However, in this domain the dictionary that contains the possible senses of every target word can be automatically extracted from the UMLS database itself. In particular, in the considered evaluation framework, which will be explained later on, this dictionary of senses is provided within the test dataset.

Given a test instance, we need to convert the plain text written in English 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. That is, the disambiguation service provided by Metamap is not used in this step, since it would give us a priori information about the possible senses of the concepts in the context of the target word. The rest of the configuration parameters are also set to their default values. The information needed to feed the graph for performing the disambiguation should be composed by all the possible types of concepts that can be found in the co-occurrence graph, that is, CUIs, nouns and adjectives in English, and nouns and adjectives in any additional language used for enriching the graph. For extracting this information, we need to obtain a translation of the test instance, which is only written in English. This translation is automatically obtained through the use of the Yandex translator,[1] an automatic translation engine which allows users to obtain translations between a large number of languages. Once we have this translation of the test instance, we can run the TreeTagger tool over the English and the translated version and select the nouns and adjectives of both texts, to enrich the original context containing only the CUIs.

In this work we are going to explore two different algorithms for disambiguation:

- **One-Step algorithm**: The first disambiguation algorithm makes use of the direct links in the graph for determining the weight of the relationships between a particular solution (one of the senses of the target term) and the concepts found in the context of the target term. For the purposes of this algorithm, we have developed a way to measure the importance of a link between two concepts of the graph, considering the significance of their co-occurrence. Given Eq. (5) in Section 3.2, 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 Eq. (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.
  Using the weighted co-occurrence graph we can rank the possible senses of the target term in the co-occurrence graph given the test instance. Given a test instance T containing a target term t and the terms of its context C, the set of possible senses of the term is represented by $S_t = s_1, s_2, \ldots, s_n$. For each s_i, we retrieve from the graph the set of concepts $S_C = c_1, c_2, \ldots, c_m$, which contains the concepts from C that are directly connected to s_i in the graph. We define the weight of a link between a concept $c_k \in S_c$ and a sense $s_i \in S_t$ to be w_{ki}. Hence, the final weight of s_i, denoted by W_i, is computed through the following formula:
  $$W_i = \sum_{k=1}^{m} w_{ki}, \quad (6)$$
  that is, the final weight of s_i is computed by adding up the weights of links between concepts from the context and s_i itself. After computing the weights of every possible sense of the target term, the system will propose the sense with the highest rank to be the most appropriate sense for the test instance.
- **Personalized PageRank**: The second algorithm that we have selected for performing this step is the Personalized PageRank algorithm, initially introduced in [47]. This algorithm is based on the PageRank algorithm [48] which has been successfully applied to WSD tasks [49]. 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:
  $$P = cMP + (1-c)v, \quad (7)$$
  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, $c = 0.85$. 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

[1] https://translate.yandex.com.

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 concepts (CUIs, nouns or adjectives in English, and nouns or adjectives in any additional language) that appear in the context of the target term we want to disambiguate. Once that we have C (concepts from the context of the test instance), we build v as a $N \times 1$ vector whose values will be $v_i = \frac{1}{|C|}$ if node i represents a concept of the context, and 0 otherwise. After performing the Personalized PageRank algorithm, we will select the node with highest rank among those representing possible senses of the target concept.

4. Test environment

4.1. Test dataset

In this section we present the dataset that will be used to evaluate the performance of our system in all the experiments. This test dataset is the NLM-WSD corpus [50], which is composed of 50 general ambiguous terms, with 100 instances per term. Each instance is an abstract downloaded from Medline containing an ambiguous term. In the creation of the corpus, each instance was manually annotated with the CUI that represents the correct sense for the target term present in the abstract. 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").

4.2. Knowledge base

For a proper analysis of the level of improvement that can be achieved by introducing multilinguality in the knowledge base used for disambiguation (the co-occurrence graph), we need to define two types of graphs: The first type is built using English documents, that is, containing CUIs, and nouns and adjectives in English. We will refer to this type of graph as "English graph" in the rest of the paper. The second type of graph can be seen as an enrichment of the former one, and is built using English documents and documents written in the other language or languages used for adding multilinguality. Hence, this second type of graph will contain CUIs, nouns and adjectives in English, and nouns and adjectives in the other language or languages. This type of graph will be referred to as "Mixed graph" in the rest of the paper.

When we create the co-occurrence graph from a set of documents, we will generate co-occurrence links between two CUIs, but also between a CUI and a word (noun or adjective), or even between words. Therefore, the final structure of the graph (number of nodes and connections between nodes) will change, and we expect this enhanced structure of the graph to improve the accuracy of the system in the WSD task. As we briefly introduced in Section 1, the size of the knowledge base is a very important parameter in this kind of tasks. The methodology that we will follow in all the experiments will be the same: we will compare the overall accuracy achieved by both English and Mixed graphs as we increase the number of documents used for building the graph (knowledge base). This way, we will study whether Mixed graphs built with small subsets of the original multilingual corpus are able to overcome results obtained by English graphs built with larger subsets of the monolingual corpus.

4.3. Evaluation criteria

The measure that will be used for evaluating the proposed system in all the experiments will be the accuracy, which can be calculated through the following formula:

$$acc = \frac{N_C}{N_T}, \tag{8}$$

where N_C is the number of correctly disambiguated instances in the test dataset and N_T is the total number of instances to be disambiguated (in this case, 3983). This value will be then expressed as a percentage by multiplying it by 100.

5. First experiment: Elsevier bilingual corpus for rare diseases

The whole objective of this work is to analyze the possible improvements that can be achieved in a Word Sense Disambiguation task when we create a knowledge base with multilingual information. Hence, we will first need a multilingual corpus with documents written in English and at least one more language, in order to create our knowledge base (the co-occurrence graph). As we showed in Fig. 2, we transform the text documents into documents containing a list of CUIs from the UMLS database, and nouns and adjectives from all the involved languages, that is, we do not take the order of occurrence of the concepts into account. Hence, we do not need to work with parallel corpora in which text is sentence-aligned, but instead we can use comparable corpora containing original documents and their translations into the additional languages.

The multilingual comparable corpus that we have used for this first experiment is the "Elsevier Bilingual Corpus for Rare Diseases" (EBCRD), which we have developed and made publicly available.[2] It is a bilingual corpus, written in both English and Spanish, and originally created by performing a search for abstracts containing rare diseases (RD) in Ibero-American Elsevier journals whose abstracts are written in both languages and contain at least one term of the test dataset. This corpus, which contains 94,003 documents per language (for a total of 188,006 documents), will eventually become the knowledge base used for disambiguation.

Once that we have annotated all the documents in the corpus (both those written and English and in Spanish), following the steps described in Section 3.1, we are able to build our English and Mixed co-occurrence graphs.

5.1. Results

Fig. 3 shows the results obtained in this first experiment. It illustrates the behavior of the proposed system when we use English graphs and Mixed graphs for performing the disambiguation, as we increase the number of documents used for building the graph. That is, from the original corpus we take N documents containing each of the possible ambiguous terms.

We can observe the improvement achieved when we add the new corpus written in a different language, in this case Spanish, to the co-occurrence graph. Specially, we find the biggest differences when the number of documents used for creating the graph is small, for example for $N = 20$ we get an accuracy of 58.45% for the English graph and 62.57% for the Mixed graph, which represents a relative improvement of 7.05%. When graphs become big-

[2] Corpus available at nlp.uned.es/~aduque/EBCRD_public.zip.

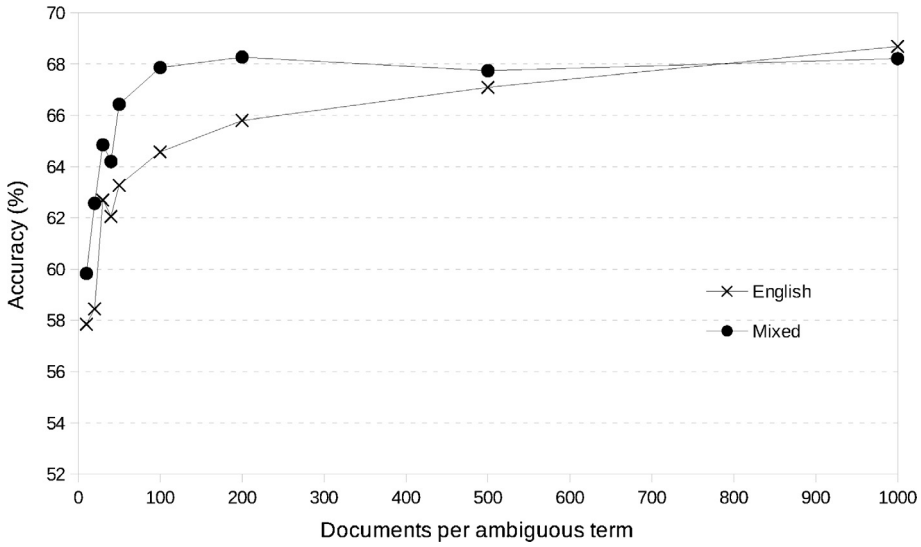

Fig. 3. Evolution of the accuracy for the English and Mixed graphs of the Elsevier corpus as we increase the number of documents per ambiguous term used for building the co-occurrence graph.

ger the improvement achieved by Mixed graphs becomes smaller. Considering what we first stated in Section 1 about the reduced size of corpora (and specifically multilingual corpora) in the biomedical domain, the fact that multilinguality performs better in smaller datasets is a good indicator.

Results shown in Fig. 3 refer to the One-Step disambiguation algorithm. We also want to compare the performance of the two considered disambiguation algorithms. Table 1 illustrates the achieved accuracy for different sizes of the set of documents used for creating the graphs, both with the One-Step and the Personalized PageRank algorithms.

In this case, we represent both the number of documents per ambiguous word and the total number of documents of the resulting graph. Since documents may contain more than one ambiguous word, the total number of documents used for building the graph is not N times the number of ambiguous words in the dataset (49), but a smaller number, as we can observe in the table. Although there is one case in which the Personalized PageRank (PPR) overcomes the results obtained by the One-Step algorithm, in general we can observe that the latter algorithm generally outperforms PPR.

5.2. Discussion

Results shown in Section 5.1 give us a first indicator of the benefits of using multilinguality when the number of documents used for building the co-occurrence graph is small. However, as we have stated before, manual translations for creating multilingual corpora are expensive and time consuming. Hence, multilingual corpora are not always available for every subset of documents, specially when the documents are very specific. This leads us to the idea of exploring automatic translations in order to generate multilingual comparable corpora that could be used in a similar way to this experiment. Once we obtain this automatically translated corpus, we will need to analyze whether the quality of the automatic translation allows the system to achieve at least a similar accuracy to the one reported in this first experiment. Table 1 has also shown that for this particular corpus, the One-Step algorithm usually performs better than the Personalized PageRank algorithm.

It is important to remark that the multilingual approach offers the best improvements when the co-occurrence graph is built with a small number of documents: between 10 and 200 documents per ambiguous term, that is, between 400 and 7000 documents in total. When the number of available documents is higher, results converge to similar accuracy values.

Table 1
Size of the document set (documents per ambiguous term and total number of documents) and comparison (accuracy in %) of the disambiguation algorithms. Bold represents the best disambiguation algorithm in each case.

Docs per term	Total # docs	One-Step	PPR
10	393	59.83	**60.61**
20	779	**62.57**	61.13
30	1185	**64.85**	62.01
40	1548	**64.20**	62.47
50	1908	**66.43**	62.87
100	3645	**67.86**	67.61
200	6853	**68.27**	64.75
500	15,202	**67.74**	62.42
1000	26,414	**68.21**	64.47

6. Second experiment: automatic translation

Considering the discussion about the first experiment, the second experiment that we propose in this work is quite straightforward: We want to analyze the performance of the system when we use automatic translations for generating a multilingual corpus, taking an English corpus as original source of information. Many different automatic translators can be found in the literature. In this case we have used the Yandex translator for generating the multilingual documents, since it provides a free API for using the translating services. It is a self-learning statistical machine translation system which creates language models and translation models through the analysis of parallel texts, and connects these models with a decoder. This decoder chooses the best option from the translation model, matches it with the language model to prove its validity, and provides statistics regarding the best result. Using this tool, we generate automatic translations from English to Spanish for every document in the Elsevier Bilingual Corpus for Rare Diseases. This way we are able to compare the performance of our system both using manual and automatic translations from the same original English corpus to enrich the knowledge base (our co-occurrence graph) with multilingual information. The

annotation step is followed in the same way as before in order to extract the CUIs and nouns and adjectives in English and Spanish that will populate the co-occurrence graph. We use the same subsets of documents for analyzing the evolution of performance as we increase the number of documents per ambiguous term.

6.1. Results

Fig. 4 completes Fig. 3 with results, for the One-Step algorithm (which performs better than PPR according to Table 1), obtained by the system with a Mixed graph created with the original English documents from the corpus, and Spanish documents created with the Yandex translator.

Results obtained by the new Mixed graph (**Mixed Yandex**) also outperforms those achieved by the English graph, and even those achieved by the original Mixed graph (**Mixed Manual**). This improvement is particularly noticeable for small subsets of documents. For example, if we consider $N = 30$ (being N the number of documents per ambiguous term), we can observe an accuracy of 62.69% for the English graph, 64.85% for the Mixed Manual graph and 65.98% for the Mixed Yandex graph. That is, the Mixed Manual graph obtains a relative improvement of 3.45% over the English graph, and the Mixed Yandex graph a relative improvement of 1.74% over the Mixed Manual graph (and 5.25% over the English graph). The Mixed Yandex graph is even able to overcome the English graph for bigger subsets of documents, for example for $N = 1000$, the Mixed Yandex graph obtains an accuracy of 69.32% and the English graph an accuracy of 68.69% (relative improvement of 0.92%). Statistical tests have been performed for testing the significance of the differences between the results offered by the English, Mixed Manual and Mixed Yandex graphs. The tests have been applied to the results obtained for the range of sizes which represents those graphs built with a small number of documents, particularly between 5 and 50 documents per word. That is the range in which we are finding the most important differences when multilingual information is added to the English graph. As the population cannot be assumed to be normally distributed in this type of tasks, we have applied the Wilcoxon Signed-Rank test [51]. The results confirm that the differences between the Mixed Manual and English graphs are statistically significant, as well as the differences between the Mixed Yandex and English graphs. However, the differences between the Mixed Yandex and Mixed Manual graphs are not statistically significant. These results indicate that the addition of multilingual information significantly improves results over the test dataset, whereas the difference between working with manual and automatic translations is not so relevant for the task.

Considering that automatic translations are far easier to obtain than manual translations, we have performed an additional experiment in which we obtain translations for a small subset of documents in other languages apart from Spanish. Table 2 shows the results obtained by the system for the English graph, and for Mixed graphs created with different combinations of languages. We have selected a subset of 50 documents per ambiguous word to analyze the results, since previous experiments have shown that multilinguality is able to get better improvements for smaller subsets of documents. The considered languages are: Spanish (SP), German (GE), Russian (RU) and Italian (IT).

We can observe results for the English graph, and Mixed graphs created with the combination of English documents and one, two, three or all of the considered additional languages. Spanish is the language that obtains better results when combined alone with English, followed by Italian, while Russian and German offer less

Table 2
Results (accuracy in %, 50 documents per ambiguous word) obtained with combination of different languages: English (EN), Spanish (SP), German (GE), Russian (RU) and Italian (IT). Bold highlights the best results for the combination of English with none, 1, 2, 3 or 4 additional languages.

Language(s)	Accuracy (%)
EN	**63.27**
EN + SP	**66.78**
EN + GE	64.75
EN + RU	65.45
EN + IT	65.50
EN + SP + GE	**67.76**
EN + SP + IT	67.26
EN + SP + RU	67.49
EN + GE + RU	65.91
EN + GE + IT	66.11
EN + RU + IT	66.01
EN + SP + GE + RU	**67.71**
EN + SP + GE + IT	67.66
EN + SP + RU + IT	66.98
EN + GE + RU + IT	66.51
EN + SP + GE + RU + IT	**67.24**

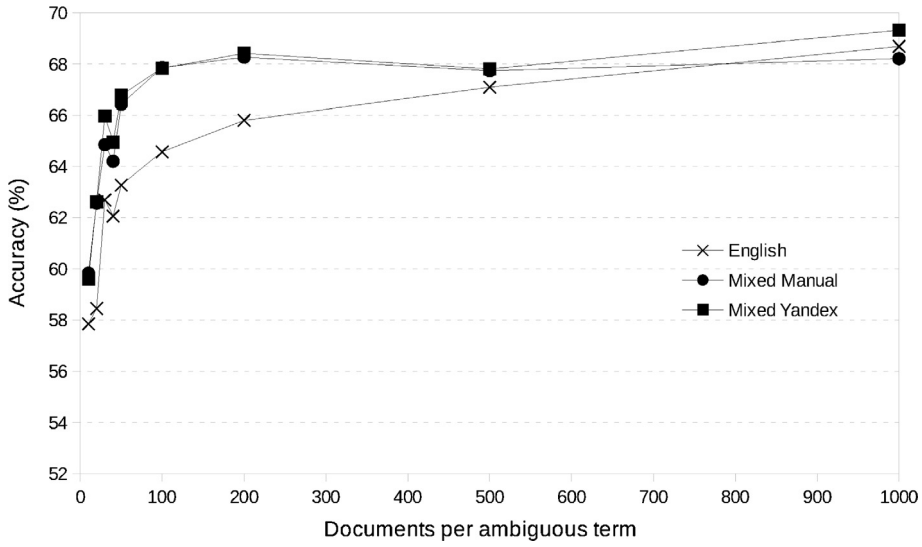

Fig. 4. Evolution of the accuracy for the English and Mixed graphs of the Elsevier corpus (**English** and **Mixed Manual**) and the Mixed graph built with its Yandex translation (**Mixed Yandex**) as we increase the number of documents per ambiguous term in the knowledge base.

improvement when combined with English. This fact may indicate that the translator is working better for Romance languages. However, the best result (accuracy of 67.76%, relative improvement of 7.10% over the English graph) is achieved when we combine Spanish and German translations with English documents. This can be due to the differences between the root languages of Spanish and German (Latin and Germanic languages respectively). The amount of information introduced by two languages of different origins and structured is probably higher than what we can expect from two similar languages, such as Spanish and Italian together. Statistical tests have also been applied in a similar way as described for Fig. 4, although we have only considered those cases interesting for the purposes of this work. In particular, we have tested the significance of the differences between results obtained only using monolingual information (row **EN** in the Table), adding multilingual information in Spanish (row **EN + SP**), since those are the conditions for previous and subsequent experiments, and adding multilingual information in Spanish and German (row **EN + SP + GE**), as that is the case in which the highest accuracy is achieved. The differences between multilingual graphs with English and Spanish information, and monolingual graphs are statistically significant, as well as those between multilingual graphs with English, Spanish and German information, and monolingual graphs. Finally, there is no statistical evidence of significant differences between using only the Spanish language for enhancing the graph, or using both Spanish and German.

6.2. Discussion

Although current machine translation systems are not able of outperforming manual translations, in our case we observe that the results obtained with a multilingual knowledge base automatically created are better than those results obtained with manually translated documents. The main reason why this could be happening is the nature of the disambiguation system described in this work and the WSD task we are facing. Apart from CUIs and nouns and adjectives from the original English documents, we are only using nouns and adjectives from the translated texts for building our knowledge base. We consider that it is likely that our system is giving more importance to the correct translation of these words than to the structure of the translated sentences (which is far more difficult for an automatic translator to represent correctly). Moreover, sometimes manual translations rely on the personal interpretation of the human translator, which can lead to less literal translations than those obtained by an automatic system. Although this can be positive when we expect a more thorough translation, in this case the creation of more literal translations can be beneficial for our purposes, since they are more likely to directly solve the ambiguity of some words. In this second experiment, we ascertain that using a small number of documents (between 10 and 200) still offers the best improvements when it comes to a multilingual approach, in this case obtained through automatic translations.

Besides, Table 2 also indicates that generating automatic Spanish translations with Yandex can offer successful results, while the combination of other languages may only slightly improve the results when the languages are different enough to provide new information.

7. Third experiment: NLM corpus

In the previous experiments we have proven the usefulness of applying multilinguality for performing WSD in the biomedical domain. However, as we stated in Section 5, the multilingual corpus used in those experiments was created from a search related to rare diseases. The information obtained in the second experiment regarding the possibility of using automatic translations allows us to explore more specific corpus, a priori written only in English, which could offer better results. In particular, as the NLM test dataset is generated from PubMed[3] abstracts, we are interested in creating a new corpus with abstracts from PubMed which contain ambiguous terms from the NLM test dataset. This way, we expect our knowledge base (the co-occurrence graph) to get closer to the characteristics of the test dataset, and hence achieve better accuracy. The last step will be to analyze whether a graph enriched with automatic translations of this "NLM-related" corpus is able to improve the accuracy of the system in the proposed WSD task, in a similar way to the previous experiments.

For this purpose, we downloaded our own set of abstracts from Medline, using the Entrez interface [52]. We performed a search for each ambiguous term of the test dataset, restricting the results to 1000 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 1000 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]

After creating the English corpus, we applied the same procedure explained in Section 6, using Yandex translator for generating the Spanish translation of each document in the English corpus. Then we perform the annotation step for extracting CUIs and nouns and adjectives in both languages, and we create the final co-occurrence English and Mixed graphs. As we did in the previous experiments, we are going to analyze the performance of the system as we increase the total number of documents used for building the graphs. In this case, the number of documents per ambiguous term is already balanced, and hence a simple random subsampling of the full corpus should be enough to obtain subsets of documents in which we find a similar number of documents per ambiguous term.

7.1. Results

Table 3 shows the results obtained by the English and Mixed graphs built with documents from the NLM-related corpus. Performance by both One-Step and PPR algorithm is also shown, to analyze whether they behave differently when graphs are built from this new corpus.

Overall accuracy obtained in this experiment is quite higher even for English graphs, probably due to the similarities between the test dataset and the NLM-related corpus used for building the graphs. Both the test dataset and the co-occurrence graph are created with abstracts downloaded from PubMed, and hence knowledge in the co-occurrence graph is more likely to present the same characteristics as the test dataset, which may lead to better results in the disambiguation process. Despite this improvement of the general results, we can observe in the table that Mixed graphs are still able to overcome English graphs, although the differences are smaller. This differences are also more important when we consider small subsets of documents (relative improvement of 2.02% for graphs built from 1000 documents), while English and Mixed graphs perform similarly when we use the complete set of 35,282 documents for building them. Differences between the disambiguation algorithms are also bigger as

[3] http://www.ncbi.nlm.nih.gov/pubmed

[4] Corpus available at nlp.uned.es/~aduque/NLM_related_public.zip.

Table 3
Results (accuracy in %) using the NLM-related corpus, for different sizes of the document set used for building the graph. Bold highlights the best configuration (algorithm and type of graph) for each experiment.

Total # docs	One-step		PPR	
	English	Mixed	English	Mixed
1 K	73.14	**74.62**	66.28	66.98
10 K	74.42	**74.67**	73.61	74.52
20 K	75.55	76.53	74.77	**76.90**
Full	76.05	77.48	**77.68**	77.63

the subset of documents is smaller, specially for 1000 documents (relative improvement of 11.41% for Mixed graphs). However, as the number of documents increases both algorithms also present similar results.

7.2. Comparative

Finally, we want to compare the best performance achieved by our multilingual system with results offered by other state-of-the-art unsupervised systems performing WSD in the biomedical domain. For this comparison, we take from Table 3 the best accuracy obtained by a Mixed graph which still present differences with the English graph of the same subset of documents. In this case, the only additional language used for the Mixed graph is Spanish, that is, concepts in the graph are CUIs, nouns and adjectives in English, and nouns and adjectives in Spanish. This best result of 76.90% of accuracy is achieved by a Mixed graph created with 20 K documents from the NLM-related corpus, selecting PPR as disambiguation algorithm, and 76.53% of accuracy under the same conditions, selecting One-Step as disambiguation algorithm (see Table 4).

In the first row, 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, these results are quite low in comparison with the accuracy achieved by our system in all the experiments reported in this work. Results from our system are then compared against different WSD systems, all of them monolingual, that is, they do not make use of multilingual information for enriching the available knowledge. The **PPR + UMLS** system [39] uses a graph-based similar approach, which makes use of a fixed graph built from the UMLS database. The **AEC** (Automatic Extracted Corpus) system [53] is a semi-supervised approach that automatically downloads and annotates abstracts for training a machine learning system. The **JDI** (Journal Descriptor Indexing) method [36] 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 [54,55] over the NLM corpus. This technique makes use of additional information from UMLS (extended definitions of the possible senses) for performing the disambiguation. 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.

Table 4
Comparison of the accuracy (%) achieved by state-of-the-art unsupervised systems (see text), and our multilingual co-occurrence graph-based system (CO-Graph). The first row corresponds to a baseline showing the performance of the Metamap disambiguation server over the test dataset. The asterisk in row **JDI** indicates modifications in the test dataset (see text).

System	NLM test dataset
Metamap baseline	49.13
PPR + UMLS	68.10
AEC	68.36
JDI	74.75*
MRD	63.89
2MRD	55.00
CO-GRAPH (One-Step)	76.53
CO-GRAPH (PPR)	**76.90**

8. Example of disambiguation

In this section a simplified example of how multilingual information can improve the performance of our system is presented. A particular case of disambiguation will be illustrated, by comparing the behavior of the our co-occurrence graph when we use only English documents for building the graph, and when multilingual (in this case, Spanish) information is added to the graph.

Fig. 5 shows this example divided in two parts: the top part of the figure presents a test instance which contains the target word "ultrasound", to be disambiguated. A look-up to the dictionary tells us that the two different senses (CUIs) of "ultrasound" between which our system should discriminate are "C0041618", referred to the process of using ultrasounds for diagnosing a disease, and "C0041621", referred to an ultrasound wave. Through the process of annotation of test instances described in Section 3.3, we obtain all the CUIs that represent concepts from the context of the test instance by applying the Metamap program to the text. Also, nouns and adjectives in both English and Spanish are extracted by running the TreeTagger tool over the original and translated text of the documents. This set of elements represents the input with which we will feed the co-occurrence graph.

The second part of the figure (bottom part) illustrates the differences of applying the disambiguation process using the English graph, or the Mixed Manual graph. The construction of these two types of graphs has been detailed in Sections 5 and 6. We can observe that the English graph does not classify this instance correctly, while in the Mixed Manual graph, the correct sense of "ultrasound" ("C0041618") is selected. If we have a look at the concepts from the context that are directly related to each of the possible senses of the target word, we observe that the English graph contains more concepts related to the wrong sense than to the correct one. That is the reason why in that case, the system selects the wrong CUI ("C0041621"). When we add multilingual information to the graph, the number of related concepts to both the target senses obviously increases. However, in the multilingual graph the sense that now presents more connections with concepts from the context is the correct one ("C0041618"), and hence the algorithm selects that CUI to be the proposed sense for this particular instance.

It is important to notice that in the example we are not explicitly illustrating the use of either of the two disambiguation algorithms studied in this work. In both algorithms it is important

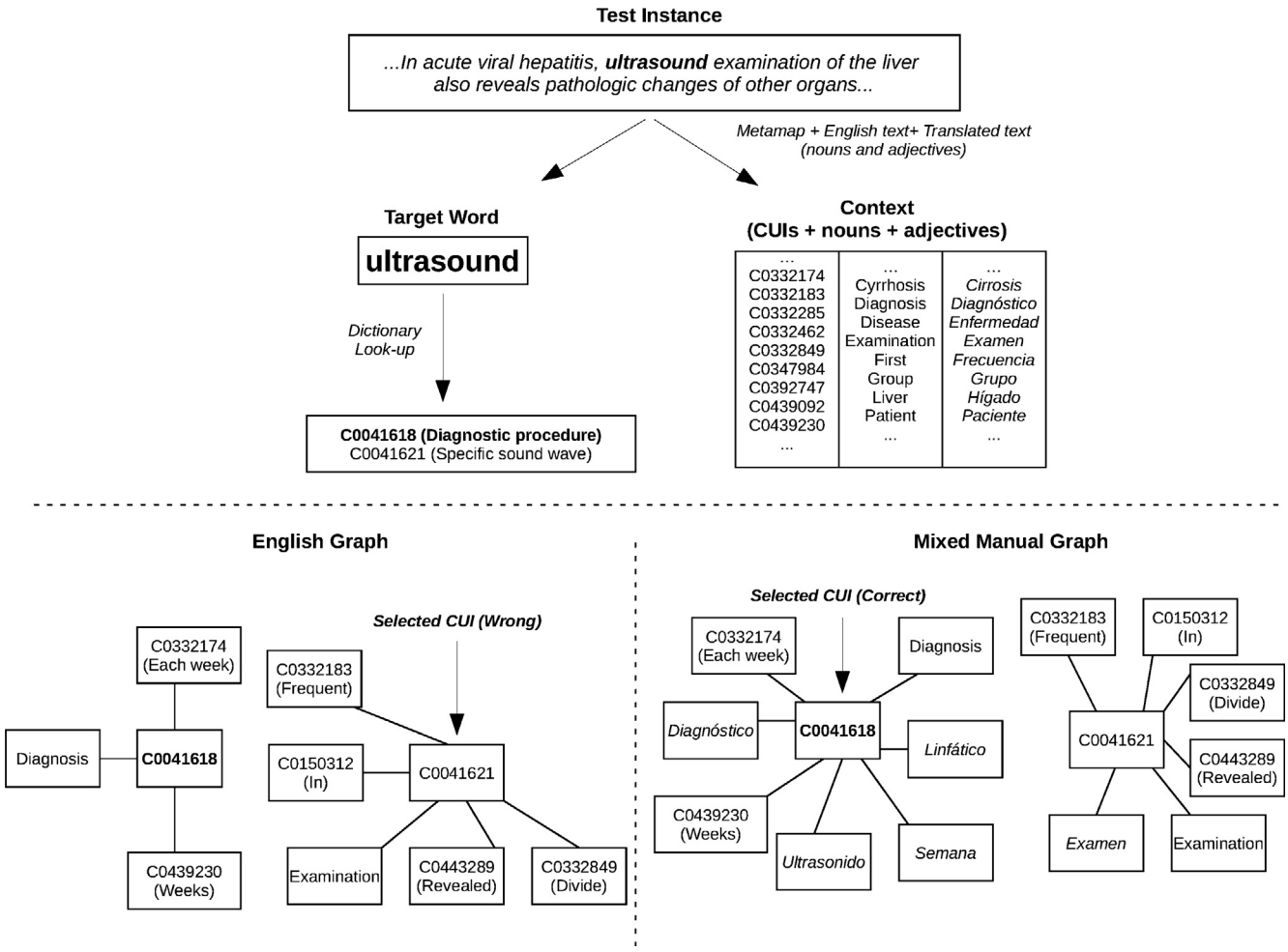

Fig. 5. Example of disambiguation of a test instance. The top part of the figure shows the annotation of the test instances, while the bottom part compares the behavior of the English graph and the Mixed Manual graph.

for a particular sense to have as many direct connections with concepts from the context as possible, in order to be selected as the most appropriate sense for an instance. Nevertheless, there exist other aspects of the algorithms that are also important, such as the weights of the links when it comes to the One-Step algorithm, or the connections between other concepts in the case of the PageRank algorithm.

9. Conclusions and future work

In this work we have presented an unsupervised system based on co-occurrence graphs for performing Word Sense Disambiguation in the biomedical domain. The objective of the study is to determine whether multilingual information is able to improve the results obtained by monolingual approaches in WSD tasks, and under which conditions this improvement is real and significative. In three different experiments performed over a test dataset widely used in the literature, multilinguality has been proven useful for WSD, particularly when the knowledge base is limited, that is, the number of documents used for building the graph is small. For the purposes of this paper, we have used a test dataset composed of general ambiguous words in biomedicine, in order to be able to analyze the performance of our system when varying the amount of available information for building the co-occurrence graph. Results obtained using two different corpora for building the graphs, one unrelated and the other related to the test dataset, indicate that a big corpus unrelated to the test dataset achieves worse results than a small corpus, but related to the test dataset (for example, the NLM-related corpus with only 1000 documents, that is, around 20 documents per ambiguous term). These facts lead us to extrapolate the results obtained in the experiments, and consider that multilinguality would also be useful when considering ambiguous words for which less occurrences could be found in the literature (for example, terms for which one of their senses represented a rare disease poorly documented).

We have observed how smaller sizes of the co-occurrence graph lead to similar or even better results than those obtained with bigger graphs, which is a very good indicator in terms of efficiency and resource consumption. For example, we can observe in Table 3 that Mixed graphs built from a subset of 1000 documents, whose size is approximately 40 K nodes and 2 million links, are able to obtain similar results to English graphs from a subset of 20,000 documents, containing 200 K nodes and more than 8 million links. The obtained improvements suggest that the translation of general terms of the context of an ambiguous term provides an important source of information to select the correct biomedical concept associated to that ambiguous biomedical term. This information can be eventually transformed into structured knowledge that allows us to disambiguate the biomedical terms in the test instances.

Automatic translations, which are normally much easier to obtain than manual translations, are able to match, and even outperform results from manual translations. This makes the

approach proposed in this work highly suitable for this kind of tasks, due to the lack of multilingual corpora in many scenarios of the biomedical domain. When using automatic translations, additional languages are proven to be more useful for WSD when their differences with the original language (in this case, English) are bigger. In general, the addition of new languages to the multilingual co-occurrence graph only improves the overall results when those languages are different enough to provide new information.

Future lines of work include the analysis of similar tasks when the original language is not English, but other languages that may present less available resources. Also, possible cross-lingual tasks for disambiguating a term written in a given language into its most suitable translation in a different target language will be explored. We also plan to apply the multilingual techniques described in this work to other tasks such as relation extraction, and in general to larger NLP systems performing more complex tasks which need WSD in their pipelines, for example, automatic text summarization.

Acknowledgments

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

References

[1] J.-A. Røttingen, S. Regmi, M. Eide, A.J. Young, R.F. Viergever, C. Årdal, J. Guzman, D. Edwards, S.A. Matlin, R.F. Terry, Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory?, The Lancet 382 (9900) (2013) 1286–1307, http://dxdoi.org/10.1016/S0140-6736(13)61046-6.

[2] S. Aymé, The importance of review articles in making the voice of rare diseases heard: Ojrd's 10th anniversary, Orph. J. Rare Dis. 11 (1) (2016) 1–2, http://dx.doi.org/10.1186/s13023-016-0456-5.

[3] C. Valmaseda, J. Martinez-Romo, L. Araujo, A tagged corpus for automatic labeling of disabilities in medical scientific papers, in: N.C.C. Chair, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France, 2016.

[4] M. Faruqui, Translation Can't Change a Name: Using Multilingual Data for Named Entity Recognition, arXiv preprint arXiv:1405.0701.

[5] J.-T. Huang, J. Li, D. Yu, L. Deng, Y. Gong, Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 7304–7308.

[6] E. Fernandez-Ordonez, R. Mihalcea, S. Hassan, Unsupervised word sense disambiguation with multilingual representations, in: LREC, 2012, pp. 847–851.

[7] E. Agirre, P. Edmonds, Word sense disambiguation: Algorithms and applications, vol. 33, Springer Science & Business Media, 2007.

[8] R. Navigli, Word sense disambiguation: a survey, ACM Comput. Surv. (CSUR) 41 (2) (2009) 10.

[9] M. Stevenson, Y. Guo, Disambiguation in the biomedical domain: the role of ambiguity type, J. Biomed. Inform. 43 (6) (2010) 972–981, http://dx.doi.org/10.1016/j.jbi.2010.08.009.

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

[11] 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, J. Biomed. Inf. 41 (6) (2008) 1088–1100, http://dx.doi.org/10.1016/j.jbi.2008.02.003.

[12] P. Resnik, D. Yarowsky, Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation, Nat. Lang. Eng. 5 (2) (1999) 113–133.

[13] M.T. Diab, P. Resnik, An unsupervised method for word sense tagging using parallel corpora, in: ACL, 2002, pp. 255–262.

[14] C. Banea, R. Mihalcea, Word sense disambiguation with multilingual features, in: Proceedings of the Ninth International Conference on Computational Semantics, IWCS '11, Association for Computational Linguistics, 2011, pp. 25–34.

[15] H. Ji, Mining name translations from comparable corpora by creating bilingual information networks, in: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora, Association for Computational Linguistics, 2009, pp. 34–37.

[16] B. Dandala, R. Mihalcea, R.C. Bunescu, Multilingual word sense disambiguation using Wikipedia, in: IJCNLP, 2013, pp. 498–506.

[17] L. Màrquez, G. Exsudero, D. Martínez, G. Rigau, Supervised corpus-based methods for WSD, in: Word Sense Disambiguation: Algorithms and Applications, Text, Speech and Language Technology, vol. 33, Springer, Dordrecht, The Netherlands, 2006, pp. 167–216.

[18] R. Mihalcea, Knowledge-based methods for WSD, in: Word Sense Disambiguation: Algorithms and Applications, Text, Speech and Language Technology, vol. 33, Springer, Dordrecht, The Netherlands, 2006, pp. 107–132.

[19] T. Pedersen, R. Bruce, A new supervised learning algorithm for word sense disambiguation, in: AAAI/IAAI, 1997, pp. 604–609.

[20] Y.K. Lee, H.T. Ng, An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, 2002, pp. 41–48.

[21] I. Iacobacci, M.T. Pilehvar, R. Navigli, Embeddings For Word Sense Disambiguation: An Evaluation Study, ACL, 2016.

[22] W. Guo, M. Diab, Coleur and colslm: a WSD approach to multilingual lexical substitution, tasks 2 and 3 semeval 2010, in: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 129–133.

[23] M. Carpuat, NRC: a machine translation approach to cross-lingual word sense disambiguation (semeval-2013 task 10), in: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 188–192.

[24] C. Fellbaum, WordNet: An Electronic Lexical Database, Bradford Books, 1998.

[25] R.S. Sinha, R. Mihalcea, Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity, ICSC, vol. 7, 2007, pp. 363–369.

[26] M. Galley, K. McKeown, Improving word sense disambiguation in lexical chaining, IJCAI, vol. 3, 2003, pp. 1486–1488.

[27] R. Mihalcea, P. Tarau, E. Figa, Pagerank on semantic networks, with application to word sense disambiguation, in: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, 2004, p. 1126.

[28] L. Plaza, M. Stevenson, A. Díaz, Resolving ambiguity in biomedical text to improve summarization, Inf. Process. Manage. 48 (4) (2012) 755–766.

[29] J.Y. Chai, A.W. Biermann, The use of word sense disambiguation in an information extraction system, in: AAAI/IAAI, 1999, pp. 850–855.

[30] M.J. Schuemie, J.A. Kors, B. Mons, Word sense disambiguation in the biomedical domain: an overview, J. Comput. Biol. 12 (5) (2005) 554–565.

[31] M. Joshi, S. Pakhomov, T. Pedersen, C.G. Chute, A comparative study of supervised learning as applied to acronym expansion in clinical reports, AMIA Annual Symposium Proceedings, vol. 2006, American Medical Informatics Association, 2006, p. 399.

[32] S. Gaudan, H. Kirsch, D. Rebholz-Schuhmann, Resolving abbreviations to their senses in medline, Bioinformatics 21 (18) (2005) 3658–3664.

[33] S. Moon, B.-T. Berster, H. Xu, T. Cohen, Word sense disambiguation of clinical abbreviations with hyperdimensional computing, AMIA Annual Symposium Proceedings, vol. 2013, American Medical Informatics Association, 2013, p. 1007.

[34] Y. Wu, J. Xu, Y. Zhang, H. Xu, Clinical abbreviation disambiguation using neural word embeddings, in: ACL-IJCNLP 2015, 2015, p. 171.

[35] M. Stevenson, Y. Guo, Disambiguation of ambiguous biomedical terms using examples generated from the UMLS metathesaurus, J. Biomed. Inform. 43 (5) (2010) 762–773.

[36] 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.

[37] B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett, The unified medical language system, J. Am. Med. Inf. Assoc. 5 (1) (1998) 1–11.

[38] 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, J. Am. Med. Inf. Assoc. 21 (5) (2014) 842–849.

[39] E. Agirre, A. Soroa, M. Stevenson, Graph-based word sense disambiguation of biomedical documents, Bioinformatics 26 (22) (2010) 2889–2896, http://dx.doi.org/10.1093/bioinformatics/btq555.

[40] 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.

[41] 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, http://dx.doi.org/10.1103/PhysRevE.84.046108.

[42] A. Duque, L. Araujo, J. Martinez-Romo, Co-graph: a new graph-based technique for cross-lingual word sense disambiguation, Nat. Lang. Eng. 21 (2015) 743–772, http://dx.doi.org/10.1017/S1351324915000091.

[43] H. Schmid, Probabilistic part-of-speech tagging using decision trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, vol. 12, 1994, pp. 44–49.

[44] A.R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program, Proc. Am. Med. Inf. Assoc. (AMIA) (2001) 17–21.

[45] D.B. Hitchcock, Yates and contingency tables: 75 years later, Journal Électronique d'Histoire des Probabilités et de la Statistique 5 (2) (2009) 1–14 (electronic only).
[46] 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, http://dx.doi.org/10.1016/j.knosys.2015.02.007.
[47] 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, http://dx.doi.org/10.1145/511446.511513.
[48] 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.
[49] 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.
[50] 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.
[51] F. Wilcoxon, Individual comparisons by ranking methods, Biomet. Bull. 1 (6) (1945) 80–83.
[52] E. Sayers, A general introduction to the e-utilities.
[53] A. Jimeno-Yepes, A.R. Aronson, Knowledge-based biomedical word sense disambiguation: comparison of approaches, BMC Bioinf. 11 (2010) 569.
[54] 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.
[55] 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 Bioinf. 12 (1) (2011) 223.