Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Co-occurrence graphs for word sense disambiguation in the biomedical domain

Andres Duque[a,c,*], 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
[c] Martinez-Romo and Lourdes Araujo: Instituto Mixto de Investigación-Escuela Nacional de Sanidad (IMIENS), Madrid 28029, Spain

ARTICLE INFO

Article history:
Received 17 July 2017
Received in revised form 23 January 2018
Accepted 11 March 2018

Keywords:
Word sense disambiguation
Graph-based systems
Unsupervised machine learning
Unified medical language system
Natural language processing
Information extraction

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.

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 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.

[*] Corresponding author.
E-mail addresses: aduque@lsi.uned.es (A. Duque), mark.stevenson@sheffield.ac.uk (M. Stevenson), juaner@lsi.uned.es (J. Martinez-Romo), lurdes@lsi.uned.es (L. Araujo).

https://doi.org/10.1016/j.artmed.2018.03.002
0933-3657/© 2018 Elsevier B.V. All rights reserved.

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 [10–13], 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 (typically, 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 calculated, 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.

Fig. 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 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 subse-

Please cite this article in press as: Duque A, et al. Co-occurrence graphs for word sense disambiguation in the biomedical domain. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.03.002