Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Knowledge based word-concept model estimation and refinement for biomedical text mining

Antonio Jimeno Yepes [a,*], Rafael Berlanga [b]

[a] Department of Computing and Information Systems, The University of Melbourne, VIC 3010, Australia
[b] Departamento de Lenguages y Sistemas Informáticos, Universitat Jaume I, Castellón de la Plana 12071, Spain

ARTICLE INFO

Article history:
Received 28 May 2014
Accepted 11 November 2014
Available online xxxx

Keywords:
Word-concept probability
Text mining
Word sense disambiguation
Information retrieval
Biomedical literature

ABSTRACT

Text mining of scientific literature has been essential for setting up large public biomedical databases, which are being widely used by the research community. In the biomedical domain, the existence of a large number of terminological resources and knowledge bases (KB) has enabled a myriad of machine learning methods for different text mining related tasks. Unfortunately, KBs have not been devised for text mining tasks but for human interpretation, thus performance of KB-based methods is usually lower when compared to supervised machine learning methods. The disadvantage of supervised methods though is they require labeled training data and therefore not useful for large scale biomedical text mining systems. KB-based methods do not have this limitation.

In this paper, we describe a novel method to generate word-concept probabilities from a KB, which can serve as a basis for several text mining tasks. This method not only takes into account the underlying patterns within the descriptions contained in the KB but also those in texts available from large unlabeled corpora such as MEDLINE. The parameters of the model have been estimated without training data. Patterns from MEDLINE have been built using MetaMap for entity recognition and related using co-occurrences.

The word-concept probabilities were evaluated on the task of word sense disambiguation (WSD). The results showed that our method obtained a higher degree of accuracy than other state-of-the-art approaches when evaluated on the MSH WSD data set. We also evaluated our method on the task of document ranking using MEDLINE citations. These results also showed an increase in performance over existing baseline retrieval approaches.

1. Introduction

Text mining of biomedical literature has supported the development of biomedical knowledge bases (KB), which are actively used by the research community [23]. These databases have contributed as well in the development of methods to perform text mining related tasks like entity recognition and relation extraction. There are a large number of KBs available for biomedical text mining purposes. Some of these resources are integrated into the Unified Medical Language System® (UMLS®) [12] and many resources are available from the Open Biological and Biomedical Ontologies (OBO) foundry [39].[1] Unfortunately, since these resources were not developed to perform text mining tasks, knowledge based methods usually exhibit lower performance compared to ad hoc supervised methods (e.g., supervised classifiers) [20]. Despite this limitation, knowledge based approaches become crucial when either there is a scarcity of labeled data to train supervised methods. Due to the heterogeneity and large scale of biomedical resources, knowledge based methods are becoming more popular.

Estimating word-concept probabilities from KBs provides an effective way to support a large range of text mining tasks in the biomedical domain [40]. Unlike supervised methods, the absence of manually labeled data can be alleviated by defining statistical approximations from either the existing data in the KBs (e.g., names, relations and descriptions) or external data such as MEDLINE® abstracts [20]. Other approaches are aimed at building statistical models directly from corpora, like Latent Dirichlet Allocation (LDA) [11], but it is not clear how to interpret or integrate these models within the KB structures [15].

Word sense disambiguation (WSD) and information retrieval (IR) are two tasks that benefit from word-concept probability models. Given an ambiguous word with its context, WSD attempts to

* Corresponding author.
E-mail address: antonio.jimeno@gmail.com (A. Jimeno Yepes).

[1] OBO foundry: http://www.obofoundry.org.

http://dx.doi.org/10.1016/j.jbi.2014.11.015
1532-0464/© 2014 Published by Elsevier Inc.

select the proper sense given a set of candidate senses. An example of ambiguity is the word *cold* which could either refer to *low temperature* or the *viral infection*. The context in which *cold* appears is used to disambiguate it. WSD is an intermediate task that supports other tasks such as: information extraction [5], information retrieval and summarization [33]. WSD in the biomedical domain is mostly based on either supervised learning or knowledge based approaches [37]. As previously mentioned, the scarcity of training data makes knowledge based methods preferable to supervised ones.

In IR, KB based methods have been proposed for either expanding queries or for performing semantic searches [14,25]. However, these methods do not provide a proper way to combine the expanded words, and just use the KB for defining improved IR queries as we have shown in [25].

This work proposes a novel method for generating word-concept statistical models from KBs that can be used directly for both IR and WSD. As mentioned earlier, this method is also able to take advantage of existing data in MEDLINE to produce a model with improved performance. These models can be integrated into IR language models to resolve ambiguity.

An implementation of the presented method is available from https://bitbucket.org/ajjimeno/wkpropability.

2. Related work

In the biomedical domain, there have been several big projects and initiatives to build comprehensive knowledge resources such as OBO and UMLS. At the same time, during the last decade researchers have devised automatic text mining techniques to find new knowledge from the scientific literature [9]. In this paper, we are interested in developing a general purpose probabilistic model that can be used in several text mining tasks, such as WSD and document ranking.

WSD methods are based on supervised learning or KB-based approaches [37]. Supervised methods are trained on examples for each one of the senses of an ambiguous word. A trained model is used to disambiguate previously unseen examples. This approach requires a large set of training examples, which is usually not available. For example, the 2009AB version of the UMLS contains approximately 24 thousand ambiguous words, based on the exact match of the words in the UMLS Metathesaurus. Preparing such training examples would be very expensive to build and maintain [44].

In the biomedical domain, KB-based methods for WSD either build a concept profile [29,28,20], develop a graph-based model [2,3] or rely on the semantic types assigned to each concept for disambiguation [19]. These derived models are compared to the context of the ambiguous word being disambiguated to select the most likely sense. In these approaches, candidate senses of the ambiguous word are UMLS concepts.

KB-based methods have been complemented with information available from existing resources like MEDLINE. An example is the use of MeSH indexing®[2] as additional information [41]; although this approach is dependent on the availability of MeSH indexing. In previous work, we collected training data from MEDLINE citations for each sense of an ambiguous word [20]. PubMed queries used to retrieve these citations were generated using English monosemous relations [27] of the candidate concepts which, potentially, have an unambiguous use in MEDLINE. This approach has shown good performance compared to other KB-based methods. In a subsequent study, we extended the work in [20] by considering all of MEDLINE instead of the top 100 recovered citations by PubMed and by generating concept profiles that can be easily estimated on large number of examples [21]. Using a large number of examples showed an improvement over previous methods.

Semi-supervised algorithms could be used to obtain additional examples of contexts for ambiguous words. We explored this in [22], where the initial disambiguation predictions provided by an unsupervised method were used as a seed to identify better concept profiles. This method showed a significant improvement.

There are several approaches in WSD that utilize the graph structure of the resources [30,1], e.g., by applying adaptations of the page rank algorithm. Unfortunately, these methods cannot be re-used for other tasks like IR, because the generated models are only able to rank senses for given contexts, and not documents for given concepts. Conversely, approaches for IR that take into account the KB (e.g., [25]) are aimed at generating IR queries but not statistical models for other purposes.

In this paper, we claim that the generation of statistical models from both the KB and existing external corpora can provide a very valuable resource for effectively performing various text mining tasks. Furthermore, we show that the presented model generates word-concept probabilities that produce good results on these tasks.

3. Methods

In this section, we present the word-concept statistical model. The estimation of the model based on the knowledge base is presented in Section 3.1. The model estimates weights to combine probabilities from concepts at different traversal steps. In this work, the model is adjusted using it for disambiguation, which is introduced in Section 3.2. The adjustment is based on Expectation–Maximization as explained in Section 3.3. Once the model is trained, it can be refined based on existing corpora in an unsupervised way as explained in Section 3.4. The word-concept probabilities obtained from this model can be used in other tasks such as IR as explained in Section 3.5. Lastly, experimental set up and data sets used in this work are presented in Section 3.6.

In this work, a KB is defined as an inventory of concepts $\mathcal{C}$, where each concept $c \in \mathcal{C}$ is associated to a list of lexical forms $lex(c)$ (i.e., strings of text that are synonyms, variants, and so on), and a set of relations to other concepts, denoted with $r(c, c')$. These relations can be of any kind, from taxonomic *is-a* relations to other specific biomedical domain relationships (e.g., treats). Resources like the UMLS Metathesaurus fit this KB definition (see Section 3.6). Strings of text consist of tokens, that are their model primitives. Tokens may be punctuation or words, which are the minimal semantic tokens in the text. Terms are words or multi-word expressions denoting a concept (e.g., the synonyms and lexical variants linked to concepts in the UMLS).

3.1. Word-concept probability estimation

We propose estimating the probability $P(w_j|c_i)$ by selecting a word w_j given a concept c_i in a KB. This is done by selecting a word from the concept c_i, step 0, or from any of the related concepts at any specific step k while traversing the KB relations. The method described below provides a way to estimate this probability at different traversal steps.

The models obtained at different steps are combined using a linear combination. The weights of the linear combination are defined in the vector $\overrightarrow{\beta}$ (from Eq. (2)), whose dimension is the number of traversal steps as shown in Eq. (1).

[2] NLM's controlled vocabulary used to index MEDLINE: https://www.nlm.nih.gov/mesh.

Q1 Please cite this article in press as: Jimeno Yepes A, Berlanga R. Knowledge based word-concept model estimation and refinement for biomedical text mining. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.11.015