Journal of Biomedical Informatics 46 (2013) 1116–1124

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text

Bridget T. McInnes [a,*], Ted Pedersen [b]

[a] Minnesota Supercomputing Institute, University of Minnesota, 117 Pleasant St SE, Minneapolis, MN 55455
[b] Department of Computer Science, University of Minnesota, 1114 Kirby Drive, Duluth, MN 55812, USA

ARTICLE INFO

Article history:
Received 27 February 2013
Accepted 17 August 2013
Available online 4 September 2013

Keywords:
Natural language processing
NLP
Word sense disambiguation
WSD
Semantic similarity and relatedness
Biomedical documents

Introduction: In this article, we evaluate a knowledge-based word sense disambiguation method that determines the intended concept associated with an ambiguous word in biomedical text using semantic similarity and relatedness measures. These measures quantify the degree of similarity or relatedness between concepts in the Unified Medical Language System (UMLS). The objective of this work is to develop a method that can disambiguate terms in biomedical text by exploiting similarity and relatedness information extracted from biomedical resources and to evaluate the efficacy of these measure on WSD.

Method: We evaluate our method on a biomedical dataset (MSH-WSD) that contains 203 ambiguous terms and acronyms.

Results: We show that information content-based measures derived from either a corpus or taxonomy obtain a higher disambiguation accuracy than path-based measures or relatedness measures on the MSH-WSD dataset.

Availability: The WSD system is open source and freely available from http://search.cpan.org/dist/UMLS-SenseRelate/. The MSH-WSD dataset is available from the National Library of Medicine http://wsd.nlm.nih.gov.

1. Introduction

Word Sense Disambiguation (WSD) is the task of automatically identifying the intended sense (or concept) of an ambiguous word based on the context in which the word is used. In our work, the set of possible meanings for a word are defined by Concept Unique Identifiers (CUIs) associated with a particular term in the Unified Medical Language System (UMLS). Thus, when performing WSD of biomedical terms, our more specific goal is to assign a term one of its possible CUIs based on its surrounding context. For example, the term *cold* could refer to the temperature (C0009264) or the common cold (C0009443), depending on the context in which it occurs.

Automatically identifying the intended concept of ambiguous words improves the performance of clinical and biomedical applications such as medical coding and indexing for quality assessment, cohort discovery and other secondary uses of data. These capabilities are becoming essential tasks due to the growing amount of information available to researchers, the transition of health care documentation towards electronic health records, and the push for quality and efficiency in health care.

The SenseRelate algorithm introduce by Patwardhan et al. [1] determines the most context-appropriate concept of an ambiguous word using the degree of semantic similarity or relatedness between the possible concepts and the terms surrounding the ambiguous word. The underlying assumption of the algorithm is that an ambiguous word will refer to the concept that is most similar to the concepts associated with the terms that surround it.

We classify semantic similarity measures in three categories: path-based measures which rely on the hierarchical relations between the terms in a taxonomy; corpus-based information content (IC) measures which augment the path information with probabilities derived from a corpus of text; and taxonomy-based IC measures which calculate the information content of a concept based on its specificity within a taxonomy. Relatedness measures use the terms found in the definitions of concepts and possibly augment those definitions with information derived from corpora. One such measure, *vector*, uses secondary co-occurrence information obtained from a corpus to determine the relatedness between terms.

In this article, we compare path-based similarity measures, corpus-based and taxonomy-based information content similarity measures, and relatedness measures. Previous studies compared

* Corresponding author.
E-mail addresses: btmcinnes@gmail.com (B.T. McInnes), tpederse@d.umn.edu (T. Pedersen).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.008

just path and corpus-based information [2] or path-based and taxonomy-based measures [3]. Overall, we found the corpus-based similarity measures perform on par or better than the taxonomy-based measures and significantly better than the path-based and relatedness measures for the task of WSD.

Section 4 describes the resources used in this work. Section 2 describes previous knowledge-based WSD methods. Section 3 describes the semantic similarity and relatedness measures used in this work. Section 5 describes our method. The data used to evaluate the method is described in Section 6. The experiments are described in Section 7 and their results in Section 8, and a comparison to previous work in Section 9. Finally, conclusions and future work are presented in Section 10.

2. Related work

Existing methods that have been proposed to automatically disambiguate words in text can be classified into three groups: supervised [4,5], unsupervised [6,7] and knowledge-based methods [8].

Supervised methods use machine learning algorithms to assign concepts to instances containing the ambiguous word. The disadvantage of these types of methods is that training data needs to be created for each target word to be disambiguated. Whether this is done manually or automatically, it is infeasible to create such data on a large scale. Knowledge-based methods do not use manually or automatically generated training data, but use information from an external knowledge source and possibly a corpus of text. Unsupervised methods use the distributional characteristics of an outside corpus and do not rely on concept information or a knowledge source. In this work, we focus on knowledge-based methods.

In the biomedical domain, Humphrey et al. [9] introduce a knowledge-based method that assigns a concept to a target word by first identifying its semantic type with the assumption that each possible concept has a distinct semantic type. A semantic type (st-) vector is created for the semantic type of each possible concept using one word terms in the UMLS that have been assigned that semantic type. A target word (tw-) vector is created using the words surrounding the target word. The cosine of the angle between the tw-vector and each of the st-vectors is calculated and the concept whose st-vector is closest to the tw-vector is assigned to the target word. The limitation of this method is that two possible concepts may have the same semantic type. For example, the term *cortices* can refer to either the cerebral cortex (C0007776) or the kidney cortex (C0022655); each with the semantic type "Body Part, Organ, or Organ Component". Analysis of the 2009 Medline data[1] shows that there are 1,072,902 terms in Medline that exist in the UMLS of which 35,013 are ambiguous and 2979 have two or more concepts with the same semantic type. This indicates that approximately 12% of the ambiguous words cannot be disambiguated using this method.

Alexopoulou et al. [10] introduce the "Closest Sense" method which calculates the average shortest distance between the semantic type of a possible concept and the semantic types each of the words surrounding the target word. This is done for each possible concept, and the concept with the shortest distance is assigned to the target word. This method also assumes that each possible concept has a distinct semantic type.

Jimeno-Yepes et al. [11] introduce a variation of the MRD method which can be seen as a variation of the Lesk algorithm [12]. In this method, a concept vector (c-vector) for each possible concept of a target word is created using the definition information from the UMLS. A target word (tw-) vector is created using the words surrounding the target word. The cosine of the angle between the tw-vector and each of the c-vectors is calculated and the concept whose c-vector is closest to the tw-vector is assigned to the target word. Rather than the vectors containing frequency scores, the frequency of the terms in the vector are normalized based on their inverted concept frequency so that terms which are repeated many times within the UMLS will have less relevance. The results of subsequent experiments conducted by Jimeno-Yepes et al. [13] compared with those conducted previously by McInnes [14] show that the inverted concept frequency significantly increases the disambiguation accuracy of the MRD method.

Jimeno-Yepes et al. [11] also introduce the AEC method, a semi-supervised approach where instances of target word are trained on automatically generate training data from Medline. Medline is manually indexed with Medical Subject Headings (MeSH) terms where each term has an associated CUI in the UMLS. Citations from Medline that contain the target word and have been annotated with one of the possible senses of the target word are extracted. These citations are used as training data into a supervised WSD algorithm. Their results show that the AEC method obtained a higher disambiguation accuracy than MRD method discussed above and the PageRank method introduced by Agirre et al. [15].

Stevenson et al. [16] introduce a modification of the PageRank algorithm called Personalized Page Rank adapted by Agirre et al. [15] for WSD. PageRank is technique for scoring the vertices according to their importance in the overall structure of a graph. In this method, a vector is constructed containing the concepts of the context words surrounding the target word. PageRank is then applied over this subgraph and the concept in the graph with maximal score is assigned to the target word. The results show that Personalized PageRank obtains a higher disambiguation accuracy than PageRank and on par with the AEC method.

Garla and Brandt [3] use the SenseRelate algorithm proposed by Patwardhan et al. [1] to evaluate path-based and taxonomy-based similarity measures. In this method, each possible concept of an ambiguous word is assigned a score by summing the similarity score between it and the terms surrounding it. The authors also evaluate obtaining the surrounding concepts using cTAKES and MetaMap finding that MetaMap performs best on biomedical text where cTAKES performs best on clinical. The results show that for biomedical text the measure taxonomy-based information content measure obtained a higher disambiguation accuracy than the path-based measures, but on clinical text the reverse was found.

3. Similarity and relatedness measures

Relatedness measures quantify the degree to which two words are associated with each other (*scissors-paper*). Similarity is a subset of relatedness and quantifies how alike two concepts are based on their location within an *is-a* hierarchy (*car-vehicle*). This section describes the similarity and relatedness measures used in this work.

3.1. Similarity measures

Existing semantic similarity measures can be categorized into two groups: path-based and information content (IC)-based. Path-based measures rely on the shortest path information, whereas IC-based measures incorporate the probability of the concept occurring in a corpus of text.

3.1.1. Path-based

Rada et al. [17] introduce the conceptual distance measure which is the length of the shortest path between two concepts ($c1$ and $c2$) in MeSH using RB/RN relations. Caviedes and Cimino

[1] http://mbr.nlm.nih.gov/index.shtml.