Journal of Biomedical Informatics 43 (2010) 762–773

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus

Mark Stevenson *, Yikun Guo

Natural Language Processing Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom

ARTICLE INFO

Article history:
Received 25 November 2009
Available online 10 June 2010

Keywords:
Natural Language Processing
NLP
Word Sense Disambiguation
WSD
Unified Medical Language System
UMLS

ABSTRACT

Researchers have access to a vast amount of information stored in textual documents and there is a pressing need for the development of automated methods to enable and improve access to this resource. Lexical ambiguity, the phenomena in which a word or phrase has more than one possible meaning, presents a significant obstacle to automated text processing. Word Sense Disambiguation (WSD) is a technology that resolves these ambiguities automatically and is an important stage in text understanding. The most accurate approaches to WSD rely on manually labeled examples but this is usually not available and is prohibitively expensive to create. This paper offers a solution to that problem by using information in the UMLS Metathesaurus to automatically generate labeled examples. Two approaches are presented. The first is an extension of existing work (Liu et al., 2002 [1]) and the second a novel approach that exploits information in the UMLS that has not been used for this purpose. The automatically generated examples are evaluated by comparing them against the manually labeled ones in the NLM-WSD data set and are found to outperform the baseline. The examples generated using the novel approach produce an improvement in WSD performance when combined with manually labeled examples.

1. Introduction

The number of documents relevant to biomedical science and related areas is growing at an ever increasing rate, making it difficult for researchers and practitioners to keep track of recent developments [2]. Automated methods for cataloguing, searching and navigating these documents would be of great benefit and it has been shown that providing access to on-line medical information improves decisions made by medical practitioners [3] and consumers [4].

However, lexical ambiguity, the phenomenon where a term (word or phrase) has more than one potential meaning, makes the automatic processing of text difficult. For example, “cold” has several possible meanings in the Unified Medical Language System (UMLS) Metathesaurus [5] including “common cold”, “cold sensation” and “cold temperature”. Ambiguous terms are common in biomedical documents. Weeber et al. [6] analyzed Medline abstracts and found that 11.7% of phrases were ambiguous relative to the UMLS Metathesaurus. The NLM Indexing Initiative [7] attempted to index biomedical journals with concepts from the UMLS Metathesaurus and concluded that lexical ambiguity was the biggest challenge in the automation of this process. An information extraction system originally designed to process radiology reports encountered problems with ambiguity when it was applied to more general biomedical texts [8]. During the development of an automated knowledge discovery system Weeber et al. [9] found that it was necessary to resolve the ambiguity in the abbreviation MG (which can mean ‘magnesium’ or ‘milligram’) in order to replicate a well-known literature-based discovery concerning the role of magnesium deficiency in migraine headaches [10].

The process of resolving lexical ambiguities, Word Sense Disambiguation (WSD), is regarded as an important part of the process of understanding natural language texts [11–13]. It is necessary for applications such as information extraction and text mining which are important in the biomedical domain for tasks such as automated knowledge discovery. Several studies have shown that the best WSD performance is obtained from systems based on supervised learning approaches [12–14]. These approaches require labeled training data, examples of ambiguous terms annotated with the correct meaning. It has also been shown that the performance of supervised approaches tends to improve with access to more labeled training data [15,16] so it is important to ensure that enough examples can be obtained to provide the best performance. However, labeled training data are often not available and the majority of existing resources contain only limited numbers of examples and do not reflect the ambiguities that occur within biomedical sciences. The only resource specific to this domain, the NLM-WSD corpus (see Section 4.1), contains 100 examples for 50

* Corresponding author.
E-mail addresses: m.stevenson@dcs.shef.ac.uk (M. Stevenson), y.guo@sheffield.ac.uk (Y. Guo).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.06.001

ambiguous domain-specific terms. But labeled training data are also extremely time-consuming and expensive to create [17,18] and it has been estimated that approximately 3.2 million sense tagged examples would be required to train a high-performance WSD system [16]. The manual labeling process is more difficult for specific domains, like biomedicine, since technical usages can only be identified by domain experts, making the process of recruiting annotators more difficult.

The costly process of manual labeling can be avoided using techniques that generate labeled examples automatically, a process that has been referred to as **pseudo-labeling** [19]. This paper describes two approaches for pseudo-labeling examples of ambiguous terms in the biomedical domain using various types of information from the UMLS Metathesaurus.

Previous approaches to pseudo-labeling are reviewed in Section 2. The two approaches based on the UMLS Metathesaurus that are used in this paper are described in Section 3. These approaches are used to generate pseudo-labeled examples for a set of 18 ambiguous terms. These examples are evaluated by using them as training data for a supervised WSD system (Section 4) and by combining them with manually labeled examples (Section 5).

2. Previous approaches to pseudo-labeling

Several approaches have been suggested for automatically generating sense tagged examples. One makes use of the fact that different senses of ambiguous words often have different translations [20,21]. For example, the word ‘‘drug” is translated to French as ‘‘médicament” when it is used to mean ‘medicine’ and ‘‘droguer” when it means ‘narcotic’. If text and its associated translation (known as ‘‘parallel text”) are available it can be used to generate sense tagged examples with the alternative translations acting as sense labels. However, the alternative translations do not always correspond to the sense distinctions in the original language and parallel text is normally difficult to obtain.

An alternative technique that does not require parallel text but relies on a lexical knowledge base has also been suggested [22]. This used WordNet [23], a lexicon that is widely used in Natural Language Processing research. The approach is based on the observation that some terms in a lexicon occur only once and, consequently, there is no doubt about their meaning. These terms are referred to as ‘‘monosemous”. However, the majority of terms have more than one possible meaning, in other words they are polysemous, and the challenge is to identify examples of the term being used with a particular meaning that can be used as training data. They suggest finding the closest related sense that is monosemous as a substitute for the ambiguous term. Sentences containing the monosemous relative are identified and the relative substituted with the ambiguous term. This approach is referred to as **monosemous relatives**. For example, the term ‘‘church” can mean ‘building’ (‘‘the *church* was empty”) and ‘institution’ (‘‘The Catholic *Church* is the largest religious body in the United States”) [24]. Monosemous relatives of the ‘building’ meaning include ‘church building’, ‘house of prayer’ and ‘synagogue’. Examples of these terms are collected and the monosemous term substituted with the polysemous one. For example, if the sentence ‘‘The synagogue is on the left at the first light” was retrieved it would be adapted to ‘‘The church is on the left at the first light” and used as an example of the ‘building’ meaning of ‘‘church”.

A variant of the monosemous relatives approach has been applied to the biomedical domain [1]. The UMLS Metathesaurus, rather than WordNet, was used to generate monosemous relatives. Terms related to the ambiguous term are identified from the Metathesaurus and unambiguous strings associated with these concepts used as the monosemous relatives. The approach was evaluated using a corpus of 35 ambiguous abbreviations from biomedical documents. They reported precision of 96.8% but recall of just 50.6%. The low recall figure indicates that the approach could only be used to generate pseudo-labeled examples for around half of the ambiguous abbreviations in the study, although the high precision score also shows that the examples that could be generated were very useful for disambiguation. There is also evidence that abbreviations are simpler to disambiguate than other ambiguous terms [25].

A variation of this approach based on semantically similar terms rather than monosemous relatives was recently proposed [19]. Terms that are semantically similar to the ambiguous word are identified using an information-theoretic algorithm for computing distributional similarity [26]. The terms identified by this process are not associated with any particular sense of the ambiguous word so an unsupervised WSD algorithm [27] is used to identify the most probable one. For example, similar terms for ‘‘church” might include ‘‘cathedral”, ‘‘chapel”, ‘‘congregation”, ‘‘parish” and ‘‘synagogue”. If we assume that the WSD algorithm identifies ‘‘cathedral”, ‘‘chapel” and ‘‘synagogue” as being related to the ‘building’ sense then examples of these terms would be identified and the relevant terms substituted with ‘‘church” to provide examples of that meaning.

A semi-supervised approach to the problem has also been applied to the biomedical domain [28]. They used techniques for Information Retrieval to analyze sense tagged examples and automatically download similar ones from Medline. They found that adding these new examples led to a small but significant improvement in the performance of their WSD system. The main problem with this approach is that it still relies on sense tagged examples. They used 100 such examples for each ambiguous term for these experiments [28].

Each of these approaches relies on external resources (parallel text, a domain ontology or an existing set of sense tagged examples). When processing biomedical text the domain ontology is the most convenient to obtain, since the UMLS is readily available. We present two pseudo-labeling approaches that used the UMLS Metathesaurus. The first of these is an extension of the ‘‘monosemous relatives” approach [1,22] and the second a novel approach that uses information about co-occurring concepts.

3. Generating sense tagged examples using the UMLS Metathesaurus

This section describes two methods for pseudo-labeling using the UMLS Metathesaurus. The first is an extension of the monosemous relatives approach (Section 3.2) and the second a novel approach that makes use of co-occurrence information (Section 3.3). Before describing the details of these approaches, the resources they make use of are described.

3.1. Resources used

3.1.1. UMLS Metathesaurus

The Unified Medical Language System (UMLS) [5] is a collection of controlled vocabularies related to biomedicine and contains a wide range of information that can be used for Natural Language Processing. The 2007AB version of the UMLS was used for the experiments described in this paper. This version was chosen since we had access to a mapping between the concepts in the UMLS and