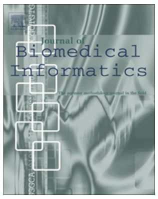
Journal of Biomedical Informatics 47 (2014) 83–90

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Determining the difficulty of Word Sense Disambiguation

Bridget T. McInnes [a],*, Mark Stevenson [b]

[a] Minnesota Supercomputing Institute, University of Minnesota, 117 Pleasant St SE, Minneapolis, MN 55455, USA
[b] 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 18 February 2013
Accepted 13 September 2013
Available online 26 September 2013

Keywords:
Natural Language Processing
NLP
Word Sense Disambiguation
WSD
Ambiguity
Biomedical documents

ABSTRACT

Automatic processing of biomedical documents is made difficult by the fact that many of the terms they contain are ambiguous. Word Sense Disambiguation (WSD) systems attempt to resolve these ambiguities and identify the correct meaning. However, the published literature on WSD systems for biomedical documents report considerable differences in performance for different terms. The development of WSD systems is often expensive with respect to acquiring the necessary training data. It would therefore be useful to be able to predict in advance which terms WSD systems are likely to perform well or badly on.

This paper explores various methods for estimating the performance of WSD systems on a wide range of ambiguous biomedical terms (including ambiguous words/phrases and abbreviations). The methods include both supervised and unsupervised approaches. The supervised approaches make use of information from labeled training data while the unsupervised ones rely on the UMLS Metathesaurus. The approaches are evaluated by comparing their predictions about how difficult disambiguation will be for ambiguous terms against the output of two WSD systems. We find the supervised methods are the best predictors of WSD difficulty, but are limited by their dependence on labeled training data. The unsupervised methods all perform well in some situations and can be applied more widely.

1. Introduction

Word Sense Disambiguation (WSD) is the task of automatically identifying the appropriate sense of an ambiguous word based on the context in which the word is used. For example, the term *cold* could refer to the *temperature* or the *common cold*, depending on how the word is used in the sentence. Automatically identifying the intended sense of ambiguous words improves the performance of biomedical and clinical applications such as medical coding and indexing; applications that are becoming essential tasks due to the growing amount of information available to researchers.

A wide range of approaches have been applied to the problem of WSD in biomedical and clinical documents [1–7]. Accurate WSD can improve the performance of biomedical text processing applications, such as summarization [8], but inaccurate WSD has been shown to reduce an application's overall performance [9]. The disambiguation of individual terms is important since some of those terms are more important than others when determining whether there is any overall improvement of the system [8]. The importance of WSD is likely to depend on the application and research question. For example, Weeber et al. [10] found that it was necessary to resolve the ambiguity in the abbreviation "MG" (which can mean "magnesium" or "milligram") in order to replicate the connection between migraine and magnesium identified by Swanson [11].

It is now possible to perform very accurate disambiguation for some types of ambiguity, such as abbreviations [12]. However, there is considerable difference in the performance of WSD systems for different ambiguities. For example, Humphrey et al. [3] report that the performance of their unsupervised WSD approach varies between 100% (for terms such as *culture* and *determination*) and 6% (for *fluid*). Consequently, it is important to determine the accuracy of a WSD system for the ambiguities of interest to get an idea of whether it will be useful for the overall application, and if so, which terms should be disambiguated.

Historically, supervised machine learning approaches have been shown to disambiguate terms with a higher degree of accuracy than unsupervised methods. The disadvantage to supervised methods is that they require manually annotated training data for each term that needs to be disambiguated. However, manual annotation is an expensive, difficult and time-consuming process which is not practical to apply on a large scale [13]. To avoid this problem, techniques for automatically labeling terms with senses have

* Corresponding author.
E-mail addresses: btmcinnes@gmail.com (B.T. McInnes), m.stevenson@dcs.shef.ac.uk (M. Stevenson).

http://dx.doi.org/10.1016/j.jbi.2013.09.009
1532-0464/© 2013 The Authors. Published by Elsevier Inc.

been developed [12,14] but these can only be applied to limited types of ambiguous terms, such as abbreviations and terms which occur with different MeSH codes. Therefore, it would be useful to be able to predict the difficulty of a particular term in order to determine whether applying WSD would be of benefit to the overall system.

This paper explores approaches to estimating the difficulty of performing WSD on ambiguities found in biomedical documents. By difficulty we mean the WSD performance that can be obtained for the ambiguity since, in practise, performance is the most important factor in determining whether applying WSD to a particular ambiguity is likely to be useful. Ambiguities for which low WSD performance is obtained are considered to be difficult to disambiguate while those for which the performance is high are considered to be easy to disambiguate.

Some of the methods applied in this paper are supervised since they are based on information derived from a corpus containing examples of the ambiguous term labeled with the correct sense. Other methods do not require this resource and only require information about the number of possible senses for each ambiguous term which is normally obtained from a knowledge source, such as the UMLS Metathesaurus (see Section 2.1.1).

Section 2 provides background information on relevant resources and techniques for computing similarity or relatedness in the biomedical domain. Section 3 describes a range of methods for estimating WSD difficulty, including ones that have been used previously and an unsupervised method based on the similarity/relatedness measures described in Section 2. Experiments to evaluate these are described in Section 4 and their results in Section 5. Finally, conclusions are presented in Section 6.

2. Resources and background

2.1. Resources

This section presents the resources that are used in the experiments described later in the paper. In particular, they are used by the similarity and relatedness measures described in Sections 2.2.1 and 2.2.2.

2.1.1. Unified Medical Language System

The Unified Medical Language System (UMLS) is a repository that stores a number of distinct biomedical and clinical resources. One such resource, used in this work, is the Metathesaurus [15].

The Metathesaurus contains biomedical and clinical concepts from over 100 disparate terminology sources that have been semi-automatically integrated into a single resource containing a wide range of biomedical and clinical information. For example, it contains the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT), which is a comprehensive clinical terminology created for the electronic exchange of clinical health information, the Foundational Model of Anatomy (FMA), which is an ontology of anatomical concepts created specifically for biomedical and clinical research, and MedlinePlus Health Topics, which is a terminology source containing health related concepts created specifically for consumers of health services.

The concepts in these sources can overlap. For example, the concept *Cold Temperature* exists in both SNOMED CT and MeSH. The Metathesaurus assigns the synonymous concepts from the various sources Concept Unique Identifiers (CUIs). Thus both the *Cold Temperature* concepts in SNOMED CT and MeSH are assigned the same CUI (C0009264). This allows multiple sources in the Metathesaurus to be treated as a single resource.

Some sources in the Metathesaurus contain additional information such as a concept's synonyms, its definition,[1] and its related concepts. The Metathesaurus contains a number of relations. The two main hierarchical relations are: the parent/child (PAR/CHD) and broader/narrower (RB/RN) relations. A parent/child relation is a hierarchical relation between two concepts that has been explicitly defined in one of the sources. For example, the concept *Cold Temperature* has an *is-a* relation with the concept *Freezing* in MeSH. This relation is carried forward to the CUI level creating a parent/child relations between the CUIs C0009264 [Cold Temperature] and C0016701 [Freezing] in the Metathesaurus. A broader/narrower relation is a hierarchical relation that does not explicitly come from a source but is created by the UMLS editors. For this work, we use the parent/child relations.

2.1.2. MEDLINE

MEDLINE[2] is a bibliographic database that currently contains over 22 million citations to journal articles in the biomedical domain and is maintained by the National Library of Medicine (NLM). The 2009 MEDLINE Baseline Repository[3] encompasses approximately 5200 journals starting from 1948 and contains 17,764,826 citations; consisting of 2,490,567 unique unigrams (single words) and 39,225,736 unique bigrams (two-word sequences). The majority of the publications are scholarly journals but a small number of other sources such as newspapers and magazines are included.

2.1.3. UMLSonMedline

UMLSonMedline, created by NLM, consists of concepts from the 2009AB UMLS and the number of times they occurred in a snapshot of MEDLINE taken on 12/01/2009. The frequency counts were obtained by using the Essie Search Engine [16] which queried MEDLINE with normalized strings from the 2009AB MRCONSO table in the UMLS. The frequency of a CUI was obtained by aggregating the frequency counts of the terms associated with the CUI to provide a rough estimate of its frequency.

2.1.4. Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) Thesaurus ([17]) is the NLM's controlled vocabulary thesaurus consisting of biomedical and health related terms/concepts created for the purpose of indexing articles from MEDLINE. Each MEDLINE citation is associated with a set of manually annotated MeSH terms that describe the content of the article. The MeSH terms are organized in a hierarchical structure in order to permit searching at various levels of specificity. The 2013 version contains 26,853 terms organized into 11 different hierarchies.[4]

2.2. Measures of similarity and relatedness

This section described measures of similarity and relatedness between biomedical concepts that have been previously explored in the literature.

2.2.1. Similarity measures

Existing semantic similarity measures can be categorized into two groups: path-based and information content (IC)-based. Path-based measures use information about the number of nodes between concepts in a hierarchy, whereas IC-based measures incorporate the probability of the concept occurring in a corpus of text.

[1] Not all concepts in the UMLS have a definition.

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

[3] http://mbr.nlm.nih.gov/.

[4] http://www.nlm.nih.gov/pubs/factsheets/mesh.html.