DALE: A Word Sense Disambiguation System for Biomedical Documents Trained using Automatically Labeled Examples

Judita Preiss and **Mark Stevenson**
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello
Sheffield S1 4DP, United Kingdom
j.preiss,m.stevenson@dcs.shef.ac.uk

Abstract

Automatic interpretation of documents is hampered by the fact that language contains terms which have multiple meanings. These ambiguities can still be found when language is restricted to a particular domain, such as biomedicine. Word Sense Disambiguation (WSD) systems attempt to resolve these ambiguities but are often only able to identify the meanings for a small set of ambiguous terms. DALE (Disambiguation using Automatically Labeled Examples) is a supervised WSD system that can disambiguate a wide range of ambiguities found in biomedical documents. DALE uses the UMLS Metathesaurus as both a sense inventory and as a source of information for automatically generating labeled training examples. DALE is able to disambiguate biomedical documents with the coverage of unsupervised approaches and accuracy of supervised methods.

1 Introduction

Word Sense Disambiguation (WSD) is an important challenge for any automatic text processing system since language contains ambiguous terms which can be difficult to interpret. Ambiguous terms that are found in biomedical documents include words, phrases and abbreviations (Schuemie et al., 2005). Identifying the correct interpretation of ambiguous terms is important to ensure that the text can be processed appropriately.

Many WSD systems developed for biomedical documents are based on supervised learning, for example (McInnes et al., 2007; Martinez and Baldwin, 2011); these have the advantage of being more accurate than unsupervised approaches. However, WSD systems based on supervised learning rely on manually labeled examples consisting of instances of an ambiguous term marked with their correct interpretations. Manually labeled examples are very expensive to create and are consequently only available for a few hundred terms, with each new domain (with its specialist vocabulary) needing new examples labeled. The majority of supervised WSD systems are limited to resolving a small number of ambiguous terms and, despite their accuracy, are not suitable for use within applications.

An alternative approach is to use *automatically labeled examples* which can be generated without manual annotation (Leacock et al., 1998). These have been used to generate an all-words WSD system that assigns senses from WordNet (Zhong and Ng, 2010). For biomedical documents the UMLS Metathesaurus (Humphreys et al., 1998b) is a more suitable lexical resource than WordNet and techniques have been developed to create automatically labeled examples for this resource (Stevenson and Guo, 2010). However, to date, automatically labeled examples have only been used as substitutes for ambiguous terms for which manually labeled examples are not available, rather than using them to create a WSD system that can resolve a wider range of ambiguities in biomedical documents.

DALE (Disambiguation using Automatically Labeled Examples) is an online WSD system for biomedical documents that was developed by creating automatically labeled examples for all ambiguous terms in the UMLS Metathesaurus. DALE is

Proceedings of the NAACL HLT 2013 Demonstration Session, pages 1–4,
Atlanta, Georgia, 10-12 June 2013.

able to identify a meaning for any term that is ambiguous in the Metathesaurus and therefore has far greater coverage of ambiguous terms than other supervised WSD systems. Other all-words WSD systems for biomedical documents are unsupervised and do not have as high accuracy as supervised approaches, e.g. (McInnes, 2008; Agirre et al., 2010). An unsupervised WSD algorithm (Humphreys et al., 1998a) is included in MetaMap (Aronson and Lang, 2010) but is unable to resolve all types of sense distinction.

2 The DALE System

2.1 Automatically Labeling Examples

DALE assigns Concept Unique Identifiers (CUIs) from the UMLS Metathesaurus. The WSD algorithm in DALE is based around a supervised algorithm (Stevenson et al., 2008) trained using automatically labeled examples. The examples are generated using two methods: *Monosemous relatives* and *Co-occurring concepts* (Stevenson and Guo, 2010). Both approaches take a single CUI, c, as input and use information from the UMLS Metathesaurus to search Medline and identify instances of c that can be used as labeled examples. The difference between the two approaches is that they make use of different information from the Metathesaurus.

Both approaches are provided with a set of ambiguous CUIs from the UMLS Metathesaurus, which represent the possible meanings of an ambiguous term, and a target number of training examples to be generated for each CUI. The UMLS Metathesaurus contains a number of data files which are exploited within these techniques, including: 1. `AMBIGLUI`: a list of cases where a LUI, a particular lexical variant of a term, is linked to multiple CUIs; 2. `MRCON`: list of all strings and concept names in the Metathesaurus; 3. `MRCOC`: co-occurring concepts.

For the monosemous relatives approach, the strings of monosemous LUIs of the target CUI and its relatives are used to search Medline to retrieve training examples. The monosemous LUIs related to a CUI are defined as any LUIs associated with the CUI in `MRCON` table and not listed in `AMBIGLUI` table. For example, one of the LUIs associated with CUI "C0028707" is L0875433 "Nutrition Science" in `MRCON` table. It is not listed in `AMBIGLUI` table and therefore considered to be a monosemous LUI of CUI "C0028707". The string "Nutrition Science" can be used to identify examples of CUI "C0028707".

The co-occurring concept approach works differently: instead of using strings of monosemous LUIs of the target CUI and its relatives, the strings associated with LUIs of a number of co-occurring CUIs of the target CUI and its relatives found in MRCOC table are used. For instance, "C0025520", "C1524024" and "C0079107" are the top three co-occurring CUIs of CUI "C0015677" in `MRCOC` table. The strings associated with LUIs of these three CUIs can be used to retrieve examples of CUI "C0015677" by searching for abstracts containing all the LUIs of the co-occurring CUIs.

These approaches were used to create labeled examples for ambiguous CUIs in the 2010AB, 2011AA, 2011AB and 2012AA versions of the UMLS Metathesaurus. Examples could be generated for 95.2%, 96.2%, 96.2% and 98% of the CUIs in each version of the Metathesaurus respectively. Neither technique was able to generate examples for the remaining CUIs, however none of these CUIs appear in the corresponding MetaMapped version of the Medline Baseline Repository (`http://mbr.nlm.nih.gov`), suggesting these CUIs do not tend to be mentioned within documents. 100 examples were generated for each CUI since using an equal number of examples for each CUI produces the best WSD performance in the absence of other information about the likelihood of each CUI (Cheng et al., 2012).

The labeled examples are converted into feature vectors consisting of lemmas of all content words in the same sentence as the ambiguous word and, in addition, the lemmas of all content words in a $\pm$4-word window around it. A single feature vector is created for each CUI by taking the centroid of the feature vectors created from the labeled examples of that CUI. These vectors are stored in the Centroid Database for later use.

2.2 Word Sense Disambiguation

WSD of an ambiguous term is carried out by compiling a list of all its possible CUIs and comparing their centroids against a feature vector created from

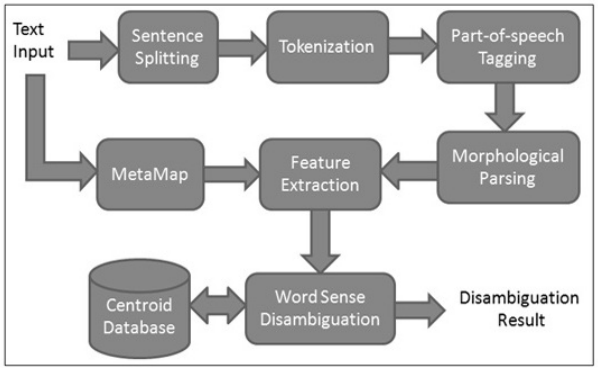

Figure 1: DALE system diagram showing the stages in the WSD process

the sentence containing the ambiguous term. Processing is carried out in multiple stages (see Fig. 1). MetaMap (Aronson and Lang, 2010) is applied to the text to identify ambiguous terms (identifying terms includes some level of multiword detection) and their possible CUIs (UMLS lookup of the identified terms). The input text is also fed into a pipeline to carry out sentence splitting, tokenization, part-of-speech tagging and morphological analysis. Information added by this pipeline is used to create a feature vector for each ambiguous term identified by MetaMap. Finally, the Word Sense Disambiguation module uses cosine similarity to compare the centroid of each possible CUI of the ambiguous term (retrieved from the Centroid Database) with the ambiguous term's feature vector (Stevenson et al., 2008). The most similar CUI is selected for each ambiguous term.

2.3 Online System

DALE is available as a web service with multiple interfaces:

The *Interactive interface* enables a user to submit a piece of text to the system and view the result in an intuitive way. Terms in the result are marked according to their polysemy: blue denotes that it has only one meaning in Metathesaurus (i.e. is not ambiguous) while green means that it has multiple meanings. Rolling the mouse over the highlighted items provides access to additional information in a tooltip style window, including the set of possible CUIs and their preferred names. Clicking on one of these CUIs links to the appropriate page from the UMLS Terminology Services (`http://uts.nlm.nih.gov/`). The CUI chosen by the WSD process is shown underlined at the bottom of the window. The result is also available in XML format which can be downloaded by clicking a link in the result page.

The *Batch interface* is more suitable for disambiguating large amounts of texts. A user can upload plain text files to be processed by DALE using the batch interface. The results will be sent to user's email address in XML format as soon as the system finishes processing the file. This interface is supported by a *Job management interface*. A job is created every time a user uploads a file and each job assigned the status of being either "Waiting" or "Running". The user is also emailed a pin code allowing them to access this interface to check the status of their jobs and cancel any waiting jobs.

3 Conclusion

This paper describes DALE, a WSD system for the biomedical domain based on automatically labeled examples. The system is able to disambiguate all ambiguous terms found in the UMLS Metathesaurus. A freely accessible web service is available and offers a set of easy to use interfaces. We intend to update DALE with new versions of the UMLS Metathesaurus as they become available.

The DALE system is available at `http://kta.rcweb.dcs.shef.ac.uk/dale/`

Acknowledgments

The authors are grateful to Weiwei Cheng for his work on the development of the original version of the DALE system. The development of DALE was funded by the UK Engineering and Physical Sciences Research Council (grants EP/H500170/1 and EP/J008427/1) and by a Google Research Award. We would also like to thank the three reviewers whose feedback has improved the clarity of this paper.

References

E. Agirre, A. Sora, and M. Stevenson. 2010. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

A. Aronson and F. Lang. 2010. An overview of MetaMap: historical perspective and recent ad-

Figure 2: Disambiguation results shown in DALE's Interactive Interface with the ambiguous term 'cold' selected. DALE shows the three possible CUIs for 'cold' identified by MetaMap with the selected CUI (C0009443) highlighted

vances. *Journal of the American Medical Association*, 17(3):229–236.

W. Cheng, J. Preiss, and M. Stevenson. 2012. Scaling up WSD with Automatically Generated Examples. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 231–239, Montréal, Canada.

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 1998a. Description of the LaSIE-II System used in MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

L. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. 1998b. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 1(5):1–11.

C. Leacock, M. Chodorow, and G. Miller. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–165.

D. Martinez and T. Baldwin. 2011. Word sense disambiguation for event trigger word detection in biomedicine. *BMC Bioinformatics*, 12(Suppl 2):S4.

B. McInnes, T. Pedersen, and J. Carlis. 2007. Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 533–537, Chicago, IL.

Bridget McInnes. 2008. An unsupervised vector approach to biomedical term disambiguation: Integrating UMLS and Medline. In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 49–54, Columbus, Ohio, June. Association for Computational Linguistics.

M. Schuemie, J. Kors, and B. Mons. 2005. Word Sense Disambiguation in the Biomedical Domain. *Journal of Computational Biology*, 12, 5:554–565.

M. Stevenson and Y. Guo. 2010. Disambiguation of Ambiguous Biomedical Terms using Examples Generated from the UMLS Metathesaurus. *Journal of Biomedical Informatics*, 43(5):762–773.

M. Stevenson, Y. Guo, R. Gaizauskas, and D. Martinez. 2008. Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics*, 9(Suppl 11):S7.

Z. Zhong and H. Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden.