AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

Integration of UMLS and MEDLINE in Unsupervised Word Sense Disambiguation

Antonio Jimeno-Yepes and **Alan R. Aronson**
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
antonio.jimeno@gmail.com, alan@nlm.nih.gov

Abstract

Scarcity of training data for word sense disambiguation argues for the use of knowledge-based disambiguation methods, which rely on information available in terminological resources. Unfortunately, these resources are not generally optimized to perform word sense disambiguation. On the other hand, there are many examples of ambiguous biomedical words with context in MEDLINE. However, these examples of ambiguity are not labeled with their proper sense. We propose the integration of the UMLS and MEDLINE to create concept profiles which are used to perform knowledge-based word sense disambiguation. Our results show an accuracy of 0.8770 on a biomedical word sense disambiguation data set; this represents a statistically significant improvement over other knowledge-based methods based on the UMLS on this data set.

Introduction

Word sense disambiguation (WSD), given an ambiguous word in context, attempts to 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 intermediary task which supports other tasks such as: information extraction (IE) (Aronson and Lang 2010), information retrieval (IR) and summarization (Plaza et al. 2011).

WSD in the biomedical domain is mostly based on supervised learning or knowledge-based (KB) approaches (Schuemie, Kors, and Mons 2005). Due to the scarcity of training data required by supervised methods, KB methods are preferred as disambiguation approaches. KB methods rely on information available in a terminological resource. On the other hand, performance of knowledge-based methods depend partly on the knowledge resource, which usually is not built to perform WSD or IR tasks (Jimeno-Yepes, Berlanga-Llavori, and Rebholz-Schuhmann 2009).

Furthermore, examples of the ambiguous words in context are available in MEDLINE, even though these ambiguous words are not labeled with the related UMLS concept. We would like to exploit the information available in MEDLINE and like many approaches, our work relies on two heuristics. The first one is one sense per discourse (Gale, Church, and Yarowsky 1992), all the occurrences of an ambiguous word in a MEDLINE citation refer to the same sense. The second one is one sense per collocation (Yarowsky 1993), so the idea is to identify the terms that tend to happen with each one of the senses of the ambiguous word. The collected examples for each on of the senses of the ambiguous words are be used to generate concept profiles. The overlap of the profiles is compared to the context of the ambiguous word. The sense of the profile with the highest overlap is selected as the predicted sense.

In this paper, we present and evaluate a method that integrates the UMLS and MEDLINE in order to provide a disambiguation model, which complements the knowledge resources. Similar to previous approaches, example citations for the sense of an ambiguous word are collected to build a disambiguation model querying MEDLINE based on the Unified Medical Language System (UMLS)®. The proposed approach builds concept profiles for a large number of concepts and is able to disambiguate terms efficiently. The top dimensions of the concept profiles denote related terms that provide relevant information to refine the information in the UMLS towards a WSD oriented version.

In the following sections, we present related work and describe the UMLS and MEDLINE. Then, we present the proposed concept profile generation approach and how it is used in disambiguation. Finally, we compare this approach to several baseline approaches and provide directions for future work.

Related work

WSD methods are based on supervised learning or knowledge-based approaches (Schuemie, Kors, and Mons 2005). 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. Knowledge-based methods rely on models built based on the information available from available knowledge sources. In the biomedical domain, this would include the UMLS. In this scenario, the candidate senses of the ambiguous word are UMLS concepts.

KB methods either build a concept profile (McInnes, Pedersen, and Carlis 2007; McInnes 2008), develop a graph-based model (Agirre, Soroa, and Stevenson 2010) in ter-

minological resources like the UMLS and work related to ontological resources (Alexopoulou et al. 2009) or rely on the semantic types assigned to each concept for disambiguation (Humphrey et al. 2006). These models are compared to the context of the ambiguous word being disambiguated. The candidate sense with highest similarity or probability is selected as the disambiguated sense.

In the biomedical domain, KB methods have been complemented with information available from existing resources like MEDLINE®. An example is the use of indexing based on MeSH®[1] (Stevenson, Agirre, and Soroa 2011) as additional information to perform disambiguation. This approach is bound to the availability of MeSH indexing. Other work, in order to collect training data for supervised methods, recover MEDLINE citations related to the different senses of an ambiguous word and train a Naïve Bayes classifier for each ambiguous word (Jimeno-Yepes and Aronson 2010). For each ambiguous word, the classes would be the candidate senses. This approach has shown a good performance compared to other KB methods. On the other hand, it requires a large number of classifiers, one per ambiguous word, which might be expensive to train and slow to use. The 2009AB version of the UMLS has 24 K ambiguous words, based on exact match of the words.

We have developed an approach that given examples collected from MEDLINE, instead of training a classifier for each ambiguous word, generates a concept profile per UMLS concept. As mentioned previously, we rely on one sense per collocation to develop the MEDLINE queries and retrieve citations related to each one of the senses of the ambiguous words. In addition, once the citation has been assigned to one of the senses given a MEDLINE query, we assume that all the mentions in the citation refer to the same sense. Each one of the generated concept profiles has as dimensions the tokens in the vocabulary and the weights are estimated based on IR techniques. We show that this approach achieves better disambiguation performance on biomedical WSD cases based and can be tuned to reduce its memory requirements.

UMLS

The NLM's UMLS (Bodenreider 2004) provides a large resource of knowledge and tools to create, process, retrieve, integrate and/or aggregate biomedical and health data. The UMLS has three main components:

- Metathesaurus®, a compendium of biomedical and health content terminological resources under a common representation which contains lexical items for each one of the concepts, relations among them and possibly one or more definitions depending on the concept. In the 2009AB version, it contains over a million concepts.
- Semantic network, which provides a categorization of Metathesaurus concepts into semantic types. In addition, it includes relations among semantic types.
- SPECIALIST lexicon, containing lexical information required for natural language processing which covers commonly occurring English words and biomedical vocabulary.

Concepts are assigned a unique identifier (CUI) which has linked to it a set of synonyms which denote alternative ways to represent the concept, for instance, in text. Concepts are assigned one or more semantic types.

MEDLINE

MEDLINE is an abbreviation for *Medical Literature Analysis and Retrieval System Online*. It is a bibliographic database begun in 1949 containing over 20 million citations to journal articles in the biomedical domain and is maintained by the National Library of Medicine (NLM). Currently, the citations are taken from approximately 5,200 journals in 37 different languages. The majority of the publications are scholarly journals but a small number of newspapers, magazines, and newsletters have been included. MEDLINE is the primary component of PubMed® [2] which is a free online repository allowing access to MEDLINE as well as other citations and abstracts in the fields of medicine, nursing, dentistry, veterinary medicine, health care systems, and pre-clinical sciences.

Concept profile preparation

The process to prepare the concept profiles in the proposed approach is split into several steps. During the first step, the ambiguous words in the UMLS are identified. Then, concept profiles are generated for each UMLS concept containing an ambiguous word. For instance, if the term *cold* is related to six concepts in the UMLS, a concept profile is created for each one of these concepts. Example citations are collected from MEDLINE for each CUI and ambiguous word pair. The citations are grouped by CUI and used to generate the concept profiles.

In the following sections this process is explained in more detail. The profiles are generated for the 2009AB version of the UMLS and a version of MEDLINE up to May 2010.

Identification of ambiguous words

To identify the ambiguous words, the first step has been to find the ambiguous words in the UMLS. We do this by examining the MRCONSO file from the UMLS. Terms in the STR field are compared between different concepts to identify ambiguous words. We have applied Porter stemming to normalize the terms and used exact matching on these terms. The outcome is the set of ambiguous words with candidate CUIs. We identified 39,820 unique ambiguous words linked to 69,508 unique CUIs. The number of ambiguous words is higher than (Jimeno-Yepes and Aronson 2010) because we used stemming to identify a larger number of potential ambiguous words. Applications like MetaMap do concept annotation allowing lexical variations that increase the ambiguity in resources like the UMLS. More background information about UMLS term normalization and ambiguity is available in (Verspoor 2005).

[1] NLM's controlled vocabulary used to index MEDLINE

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

Concept citation retrieval

The set of ambiguous words and concepts are used to build PubMed queries. Queries are generated using English monosemous relatives (Leacock, Miller, and Chodorow 1998) of the candidate concepts which, potentially, have an unambiguous use in MEDLINE. The list of candidate relatives include synonyms and terms from related concepts. In our work with the Metathesaurus, we consider a term as monosemous if it is only assigned to one concept. This means that *cold* is ambiguous since it is linked to more than one concept in the Metathesaurus while the term *cold storage* is monosemous because it is only linked to concept with CUI C0010405.

Further filtering is applied to the selected monosemous terms. Long terms (more than 50 characters) are not considered since these are unlikely to appear in MEDLINE. This avoids having unnecessarily long queries which could be problematic with retrieval systems. Very short terms (less than 3 characters) and numbers are not considered to avoid almost certain ambiguity. A standard stop word list is used to remove uninformative English terms.

The query language used by PubMed is based on Boolean operators and allows for field search, e.g. it allows searching a specific term within the metadata. Monosemous synonyms are added to the query and joined with the OR operator. Monosemous terms from related concepts are combined with the AND operator with the ambiguous term assuming one sense per collocation (Yarowsky 1993), then combined with monosemous synonyms using the OR operator. In order to retrieve documents where the text (title or abstract of the citation) contains the query terms, the [tiab] search field is used. Quotes are used to find exact mentions of the terms and increase precision.

One query per word and CUI pair are generated, a total of 94,835 queries. These queries are used to collect citations with examples for each ambiguous word. Two example queries for two candidate senses of the term *repair* are shown in figure 1.

We have run the queries using E-utilities[3], which provides a query language that is appropriate for our queries. On the other hand, E-utilities retrieves in a first step the PubMed identifiers (PMID) of the citations. Given the PMIDs, instead of retrieving them from E-utilities, we have recovered them from our own index, prepared to recover citations by sets of PMIDs. We have limited the search to the first 10,000 citations.

We are interested on concept profiles, once the citations are retrieved for each one of the CUI and word pairs, they are grouped by CUI. In this way, we build a profile per concept. This is different from similar methods which train a learning algorithm per ambiguous word given the retrieved citations.

Profile generation

Given a profile's citations, the text from the title and abstract is extracted. The text is tokenized and stemmed using the Porter stemmer, and stop words are removed. In addition, the mentions of the ambiguous words related to the concept being processed are removed since they will bias the disambiguation prediction to one of the candidate concepts. For instance, the word AA is related to *C0002520 (amino acid)* and to C0001972 *(alcoholic anonymous)* and is removed from the profiles generated for these two concepts.

[3]http://www.ncbi.nlm.nih.gov/books/NBK25500

CUI: C0374711

```
"Surgical repair"[tiab]
OR ("repair"[tiab] AND
                        ("Corneal Transplantation"[tiab]
                      OR "Corneal Transplantations"[tiab]
                      OR "Corneal Graftings"[tiab]
                      OR "Corneal Grafting"[tiab]
                      OR "Cornea Transplantations"[tiab]
...
                      OR "Repair of the Middle Ear"[tiab])
   )
```

CUI:C0043240

```
"Wound Healings"[tiab] OR "Wound Repair"[tiab]
OR ("repair"[tiab] AND
                       ("Granulation Tissues"[tiab]
                     OR "Natural regeneration"[tiab]
                     OR "Blood Clottings"[tiab]
                     OR "BLOOD COAG"[tiab]
                     OR "COAG BLOOD"[tiab]
...
                     OR "Integrin alphaIIbbeta3"[tiab])
   )
```

Figure 1: Example query for the term *repair*

The concept profiles have as dimension the tokens found in the retrieved citations D_c for a given concept c. Each dimension has assigned a weight wc_i according to the token frequency as shown in equation 1. $tf_{i,j}$ stands for the frequency of token i in the citation j. Its sum over all the tokens in the citations for concept c is used as normalization factor.

$$wc_i = \frac{\sum_{j \in D_c} tf_{i,j}}{\sum_{j \in D_c} \sum_{k \in tokens_j} tf_{k,j}} \quad (1)$$

Frequent tokens in the profiles might be simply very frequent words unrelated to the concept represented by the profile. We have experimented with IDF (Inverse Document Frequency) estimated from MEDLINE. IDF for token i is presented in equation 2, in which N is the total number of citations in MEDLINE. In the experiments with IDF, each weight wc_i is multiplied by its corresponding IDF_i.

$$IDF_i = \log \frac{N}{t_i} \quad (2)$$

Profile based disambiguation

To perform disambiguation, the candidate concept profiles C_w are compared to the words surrounding the ambiguous word cx. In this work, all the words in the citation are used for disambiguation. This setup works better in disambiguation compared to smaller context sizes (Jimeno-Yepes and Aronson 2010).

The comparison of the concept profile vector and the context vector is based on the cosine similarity as shown in equation 3. The candidate concept with the highest cosine similarity is selected as candidate concept. This approach is used with UMLS based concept profiles (Jimeno-Yepes and Aronson 2010).

$$\underset{c \in C_w}{\operatorname{argmax}} \frac{c \cdot cx}{|c||cx|} \quad (3)$$

Results

Disambiguation methods are compared using the accuracy measure on a test set built on examples of MEDLINE citations with ambiguous words. The test set has been developed automatically using MeSH indexing from MEDLINE (Jimeno-Yepes, McInnes, and Aronson 2011)[4]. This set is based on the 2009AB version of the Metathesaurus and MEDLINE up to May 2010. The Metathesaurus is screened to identify ambiguous terms which contain MeSH headings. Then, each ambiguous term and the MeSH headings linked to it are used to recover MEDLINE citations using PubMed where the term and only one of the MeSH headings co-occur. Because this initial set is noisy, we have filtered out some of the ambiguous terms to enhance precision of the set. The filtering process targeted cases where at least 15 examples are available for each sense, filtered out noisy examples and ensured that each ambiguous word has more than 1 character. The resulting set called MSH WSD consists of 106 ambiguous abbreviations, 88 ambiguous terms and 9 which are a combination of both, for a total of 203 ambiguous entities. For each ambiguous term/abbreviation, the data set contains a maximum of 100 instances per sense obtained from MEDLINE. The average number of senses per ambiguous word is 2 and the number of instances between the senses is balanced in most of the cases. This means that for this set a method predicting always the same sense would have an accuracy of 0.5.

To prepare the concept profiles based on UMLS and MEDLINE presented in this work (CPs), we have used the UMLS version 2009AB and a version of MEDLINE with citations published till May 2010. We have evaluated the proposed method, as well, with IDF (CPIDF). As baseline approaches, we have used two knowledge-based approaches which have already been tested on the data set used for our experiments (Jimeno-Yepes, McInnes, and Aronson 2011). The first one Machine Readable Dictionary (MRD) generates profiles from UMLS concepts, using terms linked to the concept and related terms. This algorithm can be seen as a relaxation of Lesk's algorithm (Lesk 1986), which is very expensive since the sense combination might be exponentially large even for a single sentence. Vasilescu et al. (Vasilescu, Langlais, and Lapalme 2004) have shown that similar or even better performance might be obtained disambiguating each ambiguous word separately. During disambiguation, the candidate concepts' profiles and the context of the ambiguous word are compared using the cosine similarity. The second one Automatic Extracted Corpus (AEC) trains a Naïve Bayes classifier for each ambiguous word, where there are as many classifiers as ambiguous words. In each classifier, the classes are the candidate senses. For each candidate sense, the training data is obtained querying MEDLINE to collect citations in an similar approach as the method proposed in this paper. The retrieved citations for each ambiguous word and CUI pair are used as examples to train the classifiers. The features are tokens extracted from these examples in a similar way as the concept profiles.

Table 1 compares the accuracy of the methods. The method proposed in this paper achieves a better performance compared to the baseline methods, which is statistically significant. Statistical significance is estimated using a randomization version of the two sample t-test (Cohen 1995). An increased performance is further obtained when the IDF value from MEDLINE is used, meaning that combining the concept profiles with the distribution of terms in MEDLINE helps discarding common but irrelevant terms. MRD relies solely on UMLS concepts and has the lowest performance, we see that even the AEC approach which already combines UMLS and MEDLINE achieves a better performance.

Method	Accuracy
MRD	0.8070
AEC	0.8383
CP	0.8705
CPIDF	0.8770

Table 1: Comparison of the accuracy of WSD methods

Discussion

As we have seen in the results section, the proposed approach performs better than the baseline methods. We find that adding information from MEDLINE improves the performance of WSD algorithms compared to the MRD approach. There are several reasons for this. The concept profiles of the proposed approach are generated per concept which might have a broader set of features compared to the features available with the AEC approach which relies on the combination of ambiguous words and candidate concepts. In addition, TF-IDF has been shown to produce better results combined with Naïve Bayes (Rennie et al. 2003) and might further cause the differences between AEC and the proposed approach.

Furthermore, the profiles summarize the citations used to generate the concept profiles. In tables 2 and 3, we show the dimensions ranked by decreasing weight for two ambiguous words and candidate concepts. These dimensions, which are made of stemmed terms obtained from the retrieved citations, might potentially contribute to producing an optimized version of the UMLS. The ambiguous acronym AA stands for *Amino Acid* with CUI C0002520 or *Alcoholics Anonymous* with CUI C0001972. The ambiguous word cement stands for *bonding substance used in restorative and orthodontic dental procedures* with CUI C1706094 or *bone like substance covering the root of the tooth* with C0011343. We find terms which usually tend to occur in MEDLINE

[4]Available from: http://wsd.nlm.nih.gov/collaboration.shtml

with the ambiguous word which might not appear in related concepts in MEDLINE. Existing work in the domain of information retrieval (Jimeno-Yepes, Berlanga-Llavori, and Rebholz-Schuhmann 2009) has been applied here to optimize existing terminological resources which might be applied to WSD.

CUI: C0002520	Weights	CUI: C0001972	Weights
amino	7.2169	alcohol	20.9273
acid	5.7039	anonym	10.0203
protein	4.2946	treatment	5.6095
sequenc	3.6551	abstin	5.2095
gene	2.6065	drink	4.5233

Table 2: AA vector top 5 dimensions

CUI: C1706094	Weights	CUI: C0011343	Weights
bond	12.9071	cementum	20.4882
lute	11.3789	periodont	10.3977
adhes	9.8194	root	9.6944
resin	9.7478	dentin	7.3273
strength	7.6228	teeth	6.8572

Table 3: Cement vector top 5 dimensions

Profile vectors have a large number of dimensions due to the large number of tokens derived from MEDLINE citations; there is a total of 3.5 million unique tokens. On the other hand, a large number of these dimensions contribute little to WSD.

We have evaluated ranking the dimensions for each concept vector and keeping the dimensions with the highest weights. Several thresholds on the number of kept dimensions have been evaluated. Experiments show that considering the features with the highest weight, memory requirements are largely reduced while there is a small decrease in performance. As a final step in the generation of the profiles, the size of the profile vectors is reduced to the top 100 dimensions. After this step, there is a total of 200 K unique tokens.

Conclusions and Future Work

The approach presented in this paper integrates the UMLS and MEDLINE to produce concept profiles which improves the performance of similar WSD approaches. In addition, this approach avoids the problem of training and dealing with a large number of models and has a more compact representation. Furthermore, the generated concept profiles can be used to obtain a version of the UMLS optimized for WSD problems. We would like to explore this possibility in the future. An implementation of this approach is being integrated into MetaMap and will be available in a future release of MetaMap and its associated web services. The work presented till now has been focused on MEDLINE, we plan to evaluate the WSD developed methods in clinical data and compare the results with existing related work (Savova et al. 2008).

References

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

Alexopoulou, D.; Andreopoulos, B.; Dietze, H.; Doms, A.; Gandon, F.; Hakenberg, J.; Khelif, K.; Schroeder, M.; and Wächter, T. 2009. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC bioinformatics* 10(1):28.

Aronson, A., and Lang, F. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17(3):229–236.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32(Database Issue):D267.

Cohen, P. R. 1995. *Empirical methods for artificial intelligence*. Cambridge, MA, USA: MIT Press.

Gale, W.; Church, K.; and Yarowsky, D. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, 233–237. Association for Computational Linguistics.

Humphrey, S.; Rogers, W.; Kilicoglu, H.; Demner-Fushman, D.; and Rindflesch, T. 2006. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology (Print)* 57(1):96.

Jimeno-Yepes, A., and Aronson, A. 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC bioinformatics* 11:565.

Jimeno-Yepes, A.; Berlanga-Llavori, R.; and Rebholz-Schuhmann, D. 2009. Ontology refinement for improved information retrieval. *Information Processing & Management*.

Jimeno-Yepes, A.; McInnes, B.; and Aronson, A. 2011. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC bioinformatics* 12(1):223.

Leacock, C.; Miller, G.; and Chodorow, M. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics* 24(1):147–165.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, 24–26. ACM.

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

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

Plaza, L.; Jimeno-Yepes, A.; Díaz, A.; and Aronson, A. 2011. Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC bioinformatics* 12(1):355.

Rennie, J.; Shih, L.; Teevan, J.; and Karger, D. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning*, 616–623.

Savova, G.; Coden, A.; Sominsky, I.; Johnson, R.; Ogren, P.; Groen, P.; and Chute, C. 2008. Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of biomedical informatics* 41(6):1088–1100.

Schuemie, M.; Kors, J.; and Mons, B. 2005. Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology* 12(5):554–565.

Stevenson, M.; Agirre, E.; and Soroa, A. 2011. Exploiting domain information for word sense disambiguation of medical documents. *Journal of the American Medical Informatics Association*.

Vasilescu, F.; Langlais, P.; and Lapalme, G. 2004. Evaluating variants of the Lesk approach for disambiguating words. In *Proceedings of the Conference of Language Resources and Evaluations (LREC 2004)*, 633–636.

Verspoor, K. 2005. Towards a semantic lexicon for biological language processing. *Comparative and functional genomics* 6(1-2):61–66.

Yarowsky, D. 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, 266–271. Association for Computational Linguistics.