Knowledge-based and Knowledge-lean Methods Combined in Unsupervised Word Sense Disambiguation

Antonio Jimeno-Yepes
National Library of Medicine
8600 Rockville Pike
Bethesda, 20894, MD, USA
antonio.jimeno@gmail.com

Alan R. Aronson
National Library of Medicine
8600 Rockville Pike
Bethesda, 20894, MD, USA
alan@nlm.nih.gov

ABSTRACT

Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, attempting to select the proper sense of ambiguous words. For instance, the word *cold* could either refer to *low temperature* or *viral infection*.

Due to the scarcity of training data, knowledge-based and knowledge-lean methods receive attention as disambiguation methods. Knowledge-based methods compare the context of the ambiguous word to the information available in a terminological resource, but their main purpose is not word sense disambiguation. Knowledge-lean unsupervised methods rely on term distributions instead of a resource enumerating the possible senses but might be inappropriate when there is a requirement to commit to a terminological resource as a catalog for candidate senses.

We present preliminary results of the combination of knowledge-based and knowledge-lean unsupervised methods which improves the performance of knowledge-based methods between 3% and 8%. The evaluation is done on a new word sense disambiguation set which is available to the community.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Natural Language Processing**]: Word sense disambiguation

General Terms

Algorithms

Keywords

Word sense disambiguation, Unsupervised learning, UMLS

1. INTRODUCTION

Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, attempting to select the proper sense of ambiguous words. For instance, the word *cold* could either refer to *low temperature* or *viral infection*.

Several methods to perform WSD from supervised to knowledge-based approaches rely on a resource enumerating words and their possible senses. Supervised methods usually achieve the best performance in disambiguation but require training data for each ambiguous word [16]. Training data is expensive, and consequently scarce. Knowledge-based methods compare the context of the ambiguous word to the information available in the resource, so do not need training data; but their main objective is not WSD and usually achieve lower performance.

Knowledge-lean unsupervised methods rely on term distributions instead of a resource enumerating the possible senses [15]. Usually, these methods first perform a sense discrimination step and then a sense labeling step. Their main advantage is portability of methods to several domains but might not be fully adequate when the disambiguation method requires compliance with senses enumerated in a terminological resource.

Our motivation is to better cover the UMLS® terminological resource. Annotation tools like MetaMap [3] and automatic indexing of the MEDLINE® bibliographic database[1] [2, 4] will profit from improved disambiguation methods. In this paper, we present preliminary results of an approach which combines two knowledge-based methods and a knowledge-lean unsupervised method based on k-means. We show that the combination of methods provides a significant improvement compared to knowledge-based methods. Finally, we highlight a test set produced for the biomedical domain which might be of interest to the community.

2. RELATED WORK

Scarcity of training data due to its cost makes unsupervised methods more appealing compared to supervised ones. We explore alternatives to supervised methods in knowledge-based and knowledge-lean methods.

Knowledge-based methods compare the overlap of information from the terminological resource to the context of an ambiguous word [13, 14]. Related methods also use the inner structure of the terminological resource to approximate the sense bias [1], while other approaches additionally use available corpora to recover context examples of the ambiguous word to train supervised learning approaches [9].

Available corpora provide examples of the ambiguous word in context, and exploiting them might improve knowledge-based methods. Stevenson [17] relied on relevance feedback to further collect examples for machine learning approaches. Jimeno and Aronson [10] have extracted collocations relevant to candidate disambiguation senses and applied them in two knowledge-based methods. In their work, a reference corpus is analyzed, and related collocations are identified and contribute to disambiguation.

Finally, knowledge-lean unsupervised methods rely on the available corpora to discriminate candidate senses and label them. LDA

[1] http://www.nlm.nih.gov/databases/databases_medline.html

IHI'12, January 28–30, 2012, Miami, Florida, USA.

(Latent Dirichlet Allocation) [6] has been used to provide better results compared to previous approaches. But identified candidate senses might not correlate with all the senses in reference terminological resources.

We propose to combine knowledge-based [9] and knowledge-lean unsupervised approaches to WSD which would profit from the existing knowledge and catalog of senses while relying on a distributional analysis given the disambiguation corpus.

3. UMLS

The NLM's UMLS [5] 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.

4. MEDLINE

MEDLINE is an abbreviation for *Medical Literature Analysis and Retrieval System Online*. It is a bibliographic database containing over 18 million citations to journal articles in the biomedical domain and is maintained by the National Library of Medicine (NLM). Currently, the citations come from approximately 5,200 journals in 37 different languages starting from 1949. 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.

5. MACHINE READABLE DICTIONARY

This knowledge-based WSD method compares the context of the ambiguous word to the information available in a knowledge source about each one of the candidate senses. This approach has been previously used in [14] in the biomedical domain.

For each candidate concept a profile is generated. The concept profile is represented in a vector space in which each dimension is one of the unique words in the profile. The words from the concept profile are obtained from the concept definition or definitions if available, synonyms, and related concepts (excluding siblings) available within the UMLS. Stop words are discarded, and Porter stemming is used to normalize the words. In addition, the word frequency is normalized based on the inverted *concept* frequency so that terms which are repeated many times within the UMLS will have less relevance.

In this machine readable dictionary approach (MRD), vectors of concept profiles linked to an ambiguous word and word contexts are compared using cosine similarity. The concept with the highest cosine similarity is selected.

6. AUTOMATIC EXTRACTED CORPUS

To overcome the scarcity of manually annotated data, corpora to train statistical learning algorithms for ambiguous terms can be automatically obtained by retrieving documents from a large corpus. We call this the Automatic Extracted Corpus (AEC) approach. Queries are generated using English *monosemous relatives* [12] of the candidate concepts available from the knowledge source. The list of candidate relatives includes synonyms and terms from related concepts obtained from the UMLS. We consider a term as monosemous if it is only assigned to one concept in the Metathesaurus.

Long terms (more than 50 characters) are not considered since these terms are unlikely to appear in MEDLINE. This avoids having unnecessarily long queries which could be problematic for retrieval systems. Very short terms (less than 3 characters) and numbers are also not considered to avoid almost certain ambiguity.

Documents retrieved using PubMed are assigned to the concept which was used to generate the query. This corpus is used to train a statistical learning algorithm; in this work we have used Naïve Bayes. Disambiguation is performed using the trained model with new disambiguation examples. 100 documents are collected from MEDLINE for each concept identifier.

7. KNOWLEDGE-BASED METHODS AND K-MEANS

K-means is a hard clustering algorithm which groups instances into only one of k clusters, where k is defined *a priori*. The output of the algorithm is k centroids, and instances are assigned to the centroid with the highest similarity. The algorithm starts with a pre-selection of k centroids which can be chosen via random selection or looking for the most distant points. Then the centroids are repositioned iteratively. Each iteration is split into two steps. In the first step, the instances are assigned to the centroid to which they have the highest similarity. In the second step, the centroids are updated according to the assignment performed in the previous step. These two steps are run until convergence occurs. In our experiments, we have considered spherical k-means [8] which is based on cosine similarity.

Knowledge-based methods are bounded by the content of the terminological resource while the context in the documents might contain additional content not covered by them. The combination of the knowledge based approaches (KB) and k-means is shown in figure 1. The knowledge-based methods are run on ambiguous examples and the prediction on each of the examples is used to estimate the initial k centroids. Each centroid is one of the candidate senses for an ambiguous word. The centroids are defined as the average of the vectors of the instances assigned by the knowledge-based method to the centroid. In our work, *UMLS concept* and *sense* are synonymous.

K-means steps are implemented as follows. In the first step, the instances are labeled with the centroid with the highest cosine similarity. From this labeling we obtain the centroid of the examples being labeled. In the second step, new centroids are estimated as the average of the instances assigned to them. Once k-means has

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

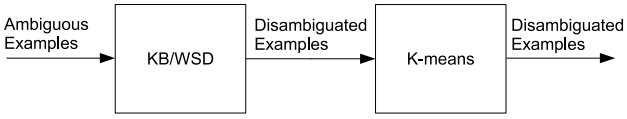

Figure 1: Combination of the KB methods with k-means

converged, the instances are labeled according to the centroid with the maximum cosine similarity.

We have evaluated the results of k-means alone as one of the baselines. In this case, the initial centroids are estimated based on some examples of disambiguated which are given to the k-means selected randomly. Several runs of the k-means are performed and the results are averaged.

8. EVALUATION

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 [11][3]. 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 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.

9. RESULTS AND DISCUSSION

Table 1 shows the accuracy of the compared methods. Statistical significance is done using a randomization version of the two sample t-test [7][4]. Disambiguation algorithms used the text provided by the title and abstract as context of the ambiguous word.

The k-means baseline uses as k the number of senses provided for each ambiguous word by the test set. The initial centroids are assigned selecting a random point from one of the senses. An average over five runs is used as the k-means baseline.

The knowledge-based disambiguation methods introduced above are compared to the combination with k-means. We have included, in addition, results extracting collocations as presented in [10], denoted by the *Coll* suffix. Extracted collocations are added to the information available to each knowledge-based method. Finally, the results are compared, as well, to Naïve Bayes with average accuracy on 10-fold cross-validation.

Of the knowledge-based methods, AEC has the highest accuracy. K-means based on random selection of centroids achieves better performance compared to the MRD approach. This result is very interesting since a small number of examples provide a result comparable to knowledge-based methods.

In addition, the combination of knowledge-based and knowledge-lean unsupervised methods improves the performance of any of these methods separately. AEC results are improved 3% while MRD results are improved 8%. Knowledge-based methods provide a reasonable initial estimate for the centroids, while the latter repositions the centroids to a more appropriate position.

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

[4]† indicates $p < 0.05$ and ‡ indicates $p < 0.01$

Method	Accuracy
K-means	0.8351
MRD	0.8070
MRD+Coll	0.8104†
MRD+KMeans	**0.8739**‡
AEC	0.8363
AEC+Coll	0.8416
AEC+KMeans	**0.8648**‡
NB	0.9386

Table 1: Results

We find as well that AEC, which has better initial performance compared to MRD, achieves a lower performance when combined with k-means. This is due to the poor performance of AEC in some of the ambiguous words compared to MRD. Problems are found mainly when dealing with acronyms related to geographical locations. An example is the acronym DE which stands for either Germany or Delaware. AEC queries, based on information from the UMLS, do not retrieve relevant documents, so the initial centroids are biased towards other groups of documents.

Compared to the the initial knowledge based methods, we find that if the candidate senses of the ambiguous have been assigned different UMLS semantic types, there is usually an improvement in the accuracy. For instance, the term *astragalus* denotes either the *astragalus plant* or the *talus bone*. In this case, the accuracy increases from 0.5152 to 0.8889. We have observed as well that a decrease in accuracy has been observed in some cases when the candidate senses have been assigned the same semantic type in the UMLS, even though is not always the case. For instance, the term *rDNA* could imply either *recombinant DNA* or *ribosomal DNA*. In this case, accuracy drops from 0.5657 to 0.3838. If we consider skipping these cases from the k-means processing, the overall accuracy is 0.8639 which is lower compared to 0.8739. This means that using k-means improves overall when the concepts are assigned the same semantic type. A decrease in performance have been found as well when the concepts belonged to different UMLS semantic types but appear in similar contexts. An example is *plague*, which either denotes a *plague* or a *plague vaccine*.

Considering previous work based on collocations, we find that the combination presented in our work improves the collocation results. The collocation analysis identified individual words which were automatically assigned to each one of the ambiguous words while our work considers all words, which builds better profiles that model better the context of ambiguous words.

Naïve Bayes still achieves the best performance even though the difference has been reduced considerably. Naïve Bayes is a supervised method, so it uses the proper sense of the ambiguous examples to build a model that performs better than purely unsupervised methods. Unsupervised methods rely on dictionary entries or distribution on the data that do not always match the senses of an ambiguous word. Supervised learning methods denote the upper bound for WSD, on the other hand is no data set available to train a system for all ambiguity cases.

10. CONCLUSIONS AND FUTURE WORK

The combination of knowledge-based and knowledge-lean unsupervised methods improves the performance of each individual method while supervised methods still achieve better performance.

We have used the knowledge-based methods to provide a proper centroid for the unsupervised approach to profit from the overlap of the information in the terminological resource and the context of the ambiguous words.

We have seen that AEC has a better disambiguation performance compared to MRD. When combined with k-means, AEC has a lower performance compared to MRD mainly due to low performance in some semantic types, e.g., related to geographical locations. We propose to combine results from AEC and MRD to produce an improved initial centroid.

In all the ambiguous cases considered in this study there was enough information available from the UMLS and MEDLINE to generate AEC and MRD models. Further study is required for the cases in which the UMLS and MEDLINE do not contain enough information.

The work presented in this paper models better the context of ambiguous words compared to previous collocation work. In our future work, we will analyze the new profiles to identify relevant clues for disambiguation.

An extension of the work will include augmenting the terminological resource with knowledge distilled from the centroids which might contribute to a better initial estimation of the starting point for k-means. Thus a resource refinement stage could be included in this method.

Finally, hard clustering of k-means might have contributed to a poor estimation of the centroids compared to soft-clustering which could improve the results. In addition, we have run the experiments on a WSD test set; we would like to extend the study to cover further examples and consider MEDLINE as the experimental corpus.

Acknowledgment

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine and by an appointment of A. Jimeno-Yepes to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

11. REFERENCES

[1] E. Agirre and A. Soroa. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics, 2009.

[2] A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, JG Mork, SJ Nelson, TC Rindflesch, and WJ Wilbur. The NLM Indexing Initiative. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2000.

[3] A.R. Aronson and F.M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229, 2010.

[4] A.R. Aronson, J.G. Mork, C.W. Gay, S.M. Humphrey, and W.J. Rogers. The NLM Indexing Initiative's Medical Text Indexer. In *Medinfo 2004: proceedings of the 11th World Conference on Medical Informatics,[San Francisco, september 7-11, 2004]*, page 268. OCSL Press, 2004.

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

[6] S. Brody and M. Lapata. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics, 2009.

[7] Paul R. Cohen. *Empirical methods for artificial intelligence*. MIT Press, Cambridge, MA, USA, 1995.

[8] I.S. Dhillon and D.S. Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1):143–175, 2001.

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

[10] A. Jimeno-Yepes and A.R. Aronson. Query Expansion for UMLS Metathesaurus Disambiguation Based on Automatic Corpus Extraction. In *Proceedings of the Ninth International Conference on Machine Learning and Applications*, 2010.

[11] A Jimeno-Yepes, BT McInnes, and AR Aronson. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223, 2011.

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

[13] M. Lesk. 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*, pages 24–26. ACM, 1986.

[14] B.T. McInnes. 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 2008. Association for Computational Linguistics.

[15] T. Pedersen. Unsupervised corpus-based methods for WSD. *Word Sense Disambiguation*, pages 133–166, 2006.

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

[17] M. Stevenson, Y. Guo, and R. Gaizauskas. Acquiring sense tagged examples using relevance feedback. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 809–816. Association for Computational Linguistics, 2008.