Query expansion for UMLS Metathesaurus disambiguation based on automatic corpus extraction

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@lhc.nlm.nih.gov

***Abstract*—Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, which attempts selecting the proper sense of ambiguous terms. In the biomedical domain, general WSD has not received much attention compared to the disambiguation of specific categories of entities like proteins and genes or diseases.**

Statistical learning approaches have achieved better performance compared to other methods. On the other hand, manually annotated data is limited, and covering all the ambiguous cases of a large resource like the UMLS is infeasible. Knowledge-based approaches using the UMLS and MEDLINE citations have achieved good performance but below that of statistical learning approaches. Our best knowledge-based result has been obtained by training a Naïve Bayes algorithm on an automatically extracted MEDLINE corpus.

In this work, we extend on previous methods to enhance the quality of an automatically extracted corpus using related terms obtained from MEDLINE without manually annotated training data. We have focused on the extraction of collocations which might be used in combination with one of the senses of the ambiguous terms. We find that left side collocations have the largest improvement in accuracy with an improvement of 4%. In addition, the combination of different types of collocations and post-filtering of retrieved citations achieves an improvement of almost 9% in accuracy.

***Keywords*-Word Sense Disambiguation; Collocation extraction; UMLS; Text categorization; Combination of approaches**

I. Introduction

Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, which attempts selecting the proper sense of ambiguous terms. In the biomedical domain, general WSD has not received much attention compared to the disambiguation of specific categories of entities like proteins and genes or diseases. We find as well some efforts devoted to acronym resolution.

Statistical learning approaches have shown better WSD performance compared to other methods. On the other hand, manually annotated data is limited and covering all the ambiguous cases of a large resource like the UMLS® is infeasible. Knowledge-based approaches [4] based on the UMLS and MEDLINE® citations have achieved good performance. Among these approaches, a method to collect training data from PUBMED®, based on automatically generated queries, which is used to train a Naïve Bayes classifier achieves better performance compared to similar Metathesaurus® based methods.

This automatically extracted corpus might lack citations for a given sense [1] or might contain false positives. We would like to improve the quality of this corpus. One way to achieve this is to improve the queries using query reformulation. Another is to filter out false positives from the retrieved set of documents.

In this paper, we take a closer look at query reformulation, extracting expansion terms from MEDLINE with the help of the UMLS Metathesaurus® and categorizing them into the proper UMLS concept.

II. Related Work

Related work already exists within information retrieval, e.g. query expansion, which could help to improve the queries built to retrieve citations for each one of the senses. For example, Stevenson et al. [8] worked on relevance feedback given some examples of disambiguated terms in context, even though a marginal improvement is obtained. In [3], left side collocations are extracted from MEDLINE to improve the performance of a knowledge-based approach based on an automatically extracted corpus. Existing work in [7] could provide a method to categorize compound nouns but this method has problems with ambiguous words; so context based WSD is proposed. In [1] the authors extract semantic information about verb's arguments which are combined with contextual features within a supervised learning environment.

We observe the following differences between ad-hoc retrieval and retrieval for an automatically extracted corpus for WSD. In the automatically extracted corpus, a candidate expansion term has to be assigned to one of the senses of the ambiguous term. In ad-hoc retrieval, the terms are assigned to the query being expanded. In addition, a candidate term

[1] In this paper, senses are denoted by UMLS concept unique identifiers (CUI).

978-0-7695-4300-0 2010

DOI 10.1109/ICMLA.2010.154

might be linked to more than one sense of the ambiguous term; so term collocation would not be effective in this case.

In the following sections we introduce several methods that we have used to identify candidate expansion terms and to assign them an ambiguous sense.

III. Methods

The methods presented in this paper are an extension of the methods published in [3], where only left side collocations are considered.

A. Left side collocations

Left side collocations are terms which act as modifiers of the ambiguous term and which occur to the left of it. Identification of left side collocations is split into collocation extraction from MEDLINE and the assignment of the collocation to one of the ambiguous senses.

Some related terms have similar semantic types but cannot be identified just by looking at a flat structure of semantic types. For instance, *cerebrospinal fluid* is assigned to *Body Substance* while the related ambiguous sense of *fluid* is assigned to *Substance*. In this work, the taxonomy of the UMLS Semantic Network is used to identify these cases. This is an improvement on [3] which relies only on a flat structure.

B. Co-occurrence collocations

Co-occurrence collocation processing has a similar structure to left side collocation. First we extract the candidate terms and then we assign them to the proper sense. On the other hand, as presented below, the implementation of the different steps are different.

1) Term extraction: Extraction of collocations from MEDLINE is performed in several steps. First, 1,000 citations are retrieved containing one of the ambiguous terms using PubMed. We have performed two experiments using these collocations. We have considered words occurring within a MEDLINE citation text and we have selected terms, on which a dependency is identified using a syntactic parser. To extract the dependent terms the citations are parsed using the Stanford Parser[2].

We determine if a term forms a collocation with the ambiguous term by comparing the probability of combined and independent events. We use the t-test as the statistical hypothesis test [6] with confidence level of $\alpha = 0.005$.

Some of these terms are not specific to one of the senses (e.g, *age*, *study*, *results*). Information retrieval literature already describes a similar problem[5]. Some of the terms are very frequent with high probability of occurrence in MEDLINE. In addition, some of them are ambiguous (e.g. *study*) [4]. We have decided to filter out terms with more than 400k occurrences in MEDLINE given as reference a standard information retrieval stop word list.

Tables I and II show examples of collocation terms.

Table I
Collocation examples based on co-occurrences

Adjustment	Determination	Repair
age	chromatography	damage
study	liquid	injury
results	standard	defect
women	chromatographic	strand
data	quantitative	excision

Table II
Collocation examples filtered using the Stanford parser

Adjustment	Determination	Repair
measures	assay	damage
illness	procedure	injury
parents	paper	dna damage
social support		techniques
		recurrence

2) Term assignment: In this step, extracted terms are assigned to one of the senses of the ambiguous term. This task is not straightforward since assigning a term to one of the ambiguous senses requires some disambiguation, and a term might be used with many of the senses. Several possibilities are available to perform the assignment of new terms:

1) The UMLS Semantic Network, which contains possible relations between semantic types, is consulted. Semantic type restrictions and text analysis could be used to determine the candidate relations. On one hand, terms not in the UMLS Metathesaurus have to be assigned a semantic type which might reduce precision. In addition, we have found that the coverage might be limited when trying to find the relation between semantic types.
2) Syntactical analysis could provide disambiguation clues; e.g selectional preferences. Available resources like PasBIO [9] for the molecular biology domain are of limited use due to the small part of the UMLS covered; we would still have to assign new terms. In addition, we do not have manually annotated sets which would allow training a system to learn extraction rules for these frames. We have performed an analysis of verbs assigned per semantic type and group but could not identify a conclusive list of verbs which could perform a reasonable categorization.
3) Yarowsky's one sense per collocation heuristic could be considered in this case. Within a collocation, the ambiguous term will have a specific sense. We could use the semantic categorization already introduced in the left side collocation approach to identify the semantic group with the highest score assigned to the citation where the candidate term and the ambiguous term co-occur, similarly to [2], to perform disambiguation. In this way, we do not need to assign a semantic

[2] http://nlp.stanford.edu/software/lex-parser.shtml

categorization to the extracted collocation term.

We have worked on the third approach since it does not require categorizing the new term into semantic types. The citation where both terms co-occur should be categorized according to semantic groups. Term collocation might help to disambiguate the sense of the term but we still have to identify the sense. On the other hand, the UMLS assigns a semantic type to each one of the concepts in the UMLS and this attribute has already been used in [2] to perform disambiguation. Assignment of the types provides an accuracy of 0.7468.

We propose to perform the assignment using a k-NN approach. To perform this categorization, 100 documents from MEDLINE are retrieved using PubMed. Then, using the group profiles presented in [3] the group with the highest score is selected. As the expansion requires high precision, we avoid taking any categorization where the number of votes is lower than 66 out of 100 votes. We have decided to choose a large number of examples and a number of neighbors over half of the examples, even though other selection could be done in the future.

IV. Results

The NLM WSD benchmark [10] is considered for the evaluation. This set contains 50 ambiguous terms and annotations of UMLS semantic types. In addition, there is a mapping to the UMLS concept unique identifiers (CUI) for the 1999 version. If there is no UMLS concept identified in the text, *None of the above* has been assigned in the NLM WSD.

We have considered the same setup as Humphrey et al.[2] and discarded the *None of the above* category. As the ambiguous term *association* has been assigned entirely to *None of the above*, it has been discarded. This means that we will present results for 49 out of the 50 ambiguous terms.

Accuracy is used to compare the approaches. Naïve Bayes is used as the statistical learning algorithm. Words occurring in the citation text, where the ambiguous terms appear, are used as the context of the ambiguous word. The corpora generated in the previous approaches are used to train this algorithm and evaluated with the NLM WSD benchmark.

In some cases, automatically generated queries have retrieved no citations for a given sense of an ambiguous term. In the experiments reported in this study we have randomly selected documents from MEDLINE for the senses in which no citation is retrieved. This has shown to improve the results for ambiguous terms like *determination* and *growth*. This also explains the difference with the results reported in [4].

Several baselines are used to compare the approaches. The first one is the Maximum Frequency Sense (MFS) baseline, where the counts are obtained from the benchmark. These frequencies are not available from any resource so no system can be built under this assumption. Results are compared as well against the machine learning trained set. This algorithm is trained and tested using the NLM WSD corpus sampled based on 10-fold cross-validation.

Results are presented in table III[3]. The original approach (Automatic) is compared against left side collocations (LSC), collocations (Coll) and collocations selected from a parse tree (CollParser). We find that left side collocations achieve better performance compared to the other methods.

In [4], we developed a filter to remove false positives based on the categorization of citations into semantic groups (Filtering). This categorization is compared to the semantic group of the UMLS concept assigned to the citation. If there is a disagreement, the citation was removed from the automatically extracted corpus. Table III shows the result of this filter. We have combined the query expansion results obtained from the LSC and CollParse and then, the retrieved set is processed by Filtering. The result shows an interesting improvement in accuracy.

Table III
Results comparing the baselines and the proposed methods

	Accuracy
Automatic	0.7017
LSC	0.7323†
Coll	0.7097
CollParser	0.7219·
Filtering	0.7265†
Combination	0.7618‡
MFS	0.8550
NB	0.8830

V. Discussion

The results presented in the previous section show that it is possible to improve over knowledge-based methods even if no manually annotated data is available.

Left side collocations seem to have the largest improvement; these collocations provide a more narrower meaning of the ambiguous term. Extracted left side collocations which can be found in the UMLS Metathesaurus are automatically classified into the proper semantic category. This means that the mistakes of the semantic group categorizer have a smaller impact. We find as well that using the UMLS Semantic Network taxonomy to link related types (e.g. Substance and Body Substance) improves over the work in [3].

Considering collocations within the citation text, we find that the performance increase is not that significant or even decreases. This might be due to categorizer mistakes. Part of these mistakes are due to terms which could either be assigned to more than a sense of the term or that are not related to any of the senses of the ambiguous terms. We have example terms like *medicine*, *practice* and *problems* assigned to one of the senses of the ambiguous sense of *pathology*.

[3] Statistical significance is tested by randomization tests. · indicates $p < 0.1$, † indicates $p < 0.05$ and ‡ indicates $p < 0.005$

Collocations restricted to dependencies with the ambiguous term seem to further filter some of the spurious terms. On the other hand, we can still see some loss in accuracy compared to the original query. For example, the term *nurse* is assigned to the ambiguous term *support*.

The approaches developed in our work rely on the ranking of categories provided by several categorizers. Different granularities should be considered in the categorization of entities because the coverage of the current approach is narrowed by the number of categories on which it can be applied. In addition, this process relies on the ranking of the categories and it considers all the text in the citation, so many different topics might be discussed in the document which might be similar to the topic of a different sense of the ambiguous term in the citation.

Finally, there are some ambiguous terms within the NLM WSD benchmark which have low performance and might be considered really polysemous, difficult to disambiguate: *blood pressure*, *pressure*, *growth*, *nutrition*. It has been more difficult to identify terms which could help to disambiguate them.

VI. Conclusions and Future Work

We have presented and evaluated several approaches to improve WSD based on an automatically retrieved corpus. Results indicate that improvement is possible without manually labeled data.

Several term extraction approaches have been studied. We have seen that left side collocations achieve the best performance. The set of extracted terms could be further extended since mostly one word collocations are extracted. More precise expansion terms could be extracted if we could identify multi-word collocations.

The assignment of semantic categories to terms and citations can be further improved. Another issue is that the granularity of the categories was either too broad (e.g. CONC in semantic groups) or too detailed (e.g. disease types). We have already rejected some of the categories; this limits the coverage of the approaches presented in this paper. A proper study of the granularity of the UMLS Semantic Network could provide better way to categorize the terms for this problem and enlarge the coverage.

The categorizer has been prepared without manually annotated training data. As the number of categories could be relatively small, manually prepared training data could improve the quality of the categorizers.

The NLM WSD data set covers a broad set of very frequent ambiguous terms in the biomedical domain. On the other hand, there are other sets of terms which could be further explored. We would like to use the knowledge-based approaches on more specific entity types like proteins and genes, or diseases.

VII. Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

[1] D. Dligach and M. Palmer. Improving Verb Sense Disambiguation with Automatically Retrieved Semantic Knowledge. In *2008 IEEE International Conference on Semantic Computing*, pages 182–189, 2008.

[2] S. Humphrey, W. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. Rindflesch. 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*, 57(1):96, 2006.

[3] A. Jimeno-Yepes and A. Aronson. Improving an automatically extracted corpus for UMLS Metathesaurus word sense disambiguation. BioSEPLN, in press.

[4] A. Jimeno-Yepes and A. Aronson. Knowledge-based biomedical word sense disambiguation: comparison of approaches. BMC Bioinformatics, in press.

[5] H. Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[6] C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 2000.

[7] B. Rosario, M. Hearst, and C. Fillmore. The descent of hierarchy, and selection in relational semantics. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 247–254. Association for Computational Linguistics, 2002.

[8] 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.

[9] T. Wattarujeekrit, P. Shah, and N. Collier. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC bioinformatics*, 5(1):155, 2004.

[10] M. Weeber, J. Mork, and A. Aronson. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMIA Symposium*, page 746. American Medical Informatics Association, 2001.