Improving an automatically extracted corpus for UMLS Metathesaurus word sense disambiguation

Mejora de un corpus extraído automáticamente para desambiguar términos del UMLS Metathesaurus

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

Resumen: Anotar a mano un conjunto de ejemplos para entrenar métodos de aprendizaje automático para desambiguar anotaciones con conceptos del UMLS Metathesaurus no es posible debido a su elevado coste. En este artículo, evaluamos dos métodos para mejorar la calidad de un corpus obtenido de manera automática. El primer método busca términos específicos y el segundo filtra falsos positivos. La combinación de los dos métodos obtiene una mejora de 6 % en F-measure y un 8 % en recall, comparado con el corpus original extraído de manera automática.

Palabras clave: Desambiguación, Extracción de terminológia, Dominio Biomédico, Estadísticas de corpus, Categorización Semántica

Abstract: Manually annotated data is expensive, so manually covering a large terminological resource like the UMLS Metathesaurus is infeasible. In this paper, we evaluate two approaches used to improve the quality of an automatically extracted corpus to train statistical learners to perform WSD. The first one contributes to more specific terms while the second filters out false positives. Using both approaches, we have obtained an improvement on the original automatic extracted corpus of approximately 6 % in F-measure and 8 % in recall.

Keywords: Word Sense Disambiguation, Term Extraction, Biomedical Domain, Corpus statistics, Semantic Categorization

1. Introduction

Word sense disambiguation (WSD) is solved efficiently when statistical learning approaches are trained on manually annotated data. Unfortunately, manual annotation is expensive, so covering a large terminological resource like the UMLS® Metathesaurus® is infeasible.

Our research interest is to provide better WSD for MetaMap (Aronson and Lang, 2010) annotation. We have compared several unsupervised disambiguation approaches (Jimeno-Yepes and Aronson, 2010) and found that statistical learning approaches trained on a corpus extracted automatically based on UMLS-built queries have better performance. This automatically extracted corpus either lacks MEDLINE® citations for some of the senses or includes false positives which need to be filtered out. In this paper, we propose to improve the quality of this corpus, which might provide an improvement on WSD trained on this corpus.

2. Related work

Some related work already exists within information retrieval, e.g. query expansion, which could help to build better queries for information retrieval. For example, (Stevenson, Guo, and Gaizauskas, 2008) worked on relevance feedback given some examples of disambiguated terms in context.

We are looking for terms to add to the query assuming Yarowsky's one sense per collocation (Yarowsky, 1995), which includes adjacent words and words neighboring the ambiguous one. Existing work by (Rosario, Hearst, and Fillmore, 2002) could provide a method to categorize compound nouns but this method has problems with ambiguous words; so context based WSD is proposed.

Approaches exist which perform citation filtering based on a given topic (Jimeno-Yepes, Berlanga-Llavori, and Rebholz-Schuhmann, 2009). We would like to explore the building of such a filters without manually annotated data.

3. Methods

In our corpus, retrieved MEDLINE citations[1] for a given sense either lack documents or include too many false positives. We propose two methods to improve the quality of an automatically built corpus to perform WSD by either further expanding the query or by filtering out potential false positives.

3.1. Query expansion

A collocation extraction process is split into two steps. In the first one, terms with high probability of forming a collocation with the ambiguous word are extracted from MEDLINE. In the second one, the terms forming a collocation are assigned, if possible, one of the senses of the ambiguous term.

3.1.1. Collocation extraction

Extraction of collocations from MEDLINE is performed in several steps. First, 1,000 citations are retrieved containing the ambiguous terms using PubMed®. Then, nouns and adjectives on the left of the ambiguous term are extracted.

We determine if a word forms a collocation with the ambiguous term comparing the probability of combined and independent events. We use the t-test as the statistical hypothesis test (Manning and Schütze, 2000) with confidence level of $\alpha = 0{,}005$. Examples are available in table 1.

Adjustment	Determination	Repair
psychosocial	quantitative	dna
psychological	spectrophotometric	excision
social	photometric	mismatch
marital	potentiometric	surgical
occlusal	accurate	hernia

Table 1: Left side collocation examples

3.1.2. Collocation categorization

Collocations extracted in the previous section have to be assigned a UMLS concept related to the senses of the ambiguous term. In refinement or adaptation of existing lexical and ontological resources, head and modifier heuristics are often used to identify new hyponyms. In our work, as the head noun is an ambiguous word, we need a different way to perform this assignment.

As each UMLS concept is assigned one or more semantic types, we propose to classify the collocations into one of these categories. We propose two ways to solve this.

The first way consists of looking for the collocation in the UMLS Metathesaurus and, if the collocation already exists, use the collocation semantic type to link it to the sense of the ambiguous term. This might be used to identify relations between existing terms in the UMLS Metathesaurus which are not related. If the collocation semantic type matches more than one of the senses of the ambiguous term, then we discard this collocation. For instance, two out of five senses of cold make reference to a distinct diseases.

The second way consists of performing the classification of collocation terms on semantic groups which are just sets of related semantic types[2]. This is done comparing a profile vector of the collocation term with a profile vector of the semantic groups of the ambiguous senses of the term using cosine similarity. Semantic group profile vector construction is explained in section 3.3.

Profile vectors for collocation terms are built by retrieving 100 citations containing the collocation from MEDLINE using PubMed. Then, the text is tokenized, words are extracted and lowercased, stop words are removed and used to build a vector with their frequency in this corpus.

If the ambiguous term has senses with the same semantic group, we do not assign any of the collocations to the senses. In addition, if any of the semantic groups is within the list of discarded semantic groups in section 3.3, then this approach is not applied.

3.2. Citation filtering

Some citations within the automatically retrieved corpus are false positives. Hence we propose a method which filters out false positives based on an automatic categorization of citations into semantic groups, similar to (Humphrey et al., 2006) with journal descriptors but based on semantic group profiles (cf. section 3.3).

We estimate the cosine similarity between the citation and the semantic groups of the concepts linked to the ambiguous terms. The group with the highest cosine similarity is compared to the one assigned in the automatic extracted corpus. If both agree, the citation is kept in the corpus and is removed otherwise.

[1] MEDLINE citations up to May 2010

[2] http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

3.3. Semantic group profiles

As we have seen in the definition of the approaches above, we can make use of categorization of terms or citations. As no manual annotation is available, we have built profile vectors for UMLS semantic categories based on MEDLINE and monosemous terms.

For each semantic type, a profile vector is built as follows. A set of monosemous terms are selected randomly from UMLS. MEDLINE citations containing these monosemous terms are retrieved using PubMed. Sentences containing the monosemous terms are selected.

This corpus is tokenized and lowercased, and stopwords are removed. Dimensions of the vector are the extracted tokens. Each dimension in the vector is assigned a weight with the frequency in the corpus multiplied by the inverted document frequency obtained from MEDLINE. As explained above, profile vectors for terms and citations are obtained in a similar way.

In table 2, top terms in the profile vectors are shown for selected semantic types. We find that semantic types *T046 (Pathologic Function)* and *T047 (Disease or Syndrome)* are quite similar; so it is difficult to provide a proper classification into semantic types given a disorder. The same thing happens with semantic types *T116 (Amino Acid, Peptide, or Protein)* and *T126 (Enzyme)*. We can map semantic types into semantic groups. In this categorization, T046 and T047 belong to the group *DISO (Disorders)* and T116 and T126 to the group *CHEM (Chemicals & Drugs)*.

Type: T046	Type: T047	Type: T116	Type: T126
patients	patients	activity	activity
management	case	delta	ec
case	hypoxic	rat	delta
cases	raeb	human	liver
diagnosis	management	liver	human

Table 2: Example top terms for profile vectors for semantic types

Semantic group profile vectors are built on the semantic type profiles. Semantic types are assigned to one or more semantic groups. Retrieved sentences belonging to a semantic type are assigned to its semantic group. This corpus is processed as explained above to produce the profile vectors. Top terms for selected semantic groups are shown in table 3.

Categories like *CONC (Concepts & Ideas)* or *ANAT (Anatomy)* do not seem to behave correctly in a manual assessment and are not considered in any of the approaches presented in this study. The CONC group is very generic and its profile seems to always rank higher than any other group profile. On the other hand, the group ANAT is never assigned since the different body parts are linked to a disorder, which is always ranked higher.

Group: DISO	Group: CHEM	Group: CONC	Group: ANAT
patients	human	health	human
case	activity	patients	rat
treatment	acid	based	cells
cases	effects	study	function
diagnosis	effect	children	anatomy

Table 3: Example top terms for profile vectors for semantic groups

4. Results

The NLM WSD benchmark (Weeber, Mork, and Aronson, 2001) is considered for the evaluation. This set contains 50 ambiguous terms and annotations of UMLS semantic types.

We have considered the same setup as Humphrey et al.(Humphrey et al., 2006) 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.

Weighted precision and recall and F-measure are 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.

Three baselines are used: the original automatic corpus (Automatic), the Maximum Frequency Sense (MFS, the counts are obtained from the benchmark) and the Naïve Bayes(NB) trained on the NLM WSD set using 10-fold cross-validation [3].

As we see in table 4, collocations (Lex. Inc.) and filtering (Filt.+Lex.) improve over the original automatic built corpus (Automatic). The largest improvement is on recall and a more modest one in precision.

The combination of the approaches proposed in this article improves over the MFS baseline in terms of F-measure but it is still far from the NB baseline. In addition, in terms of

[3] The reader can compare the results with other unsupervised techniques based on (Jimeno-Yepes and Aronson, 2010)

	Precision	Recall	F-measure
Automatic	0.8673	0.6836	0.7646
Lex. Inc.	0.8805	0.7186	0.7914
Filt.+Lex.	**0.8817**	**0.7468**	**0.8086**
MFS	0.7577	0.8550	0.8034
NB	**0.8641**	**0.8830**	**0.8735**

Table 4: Comparison of the WSD baselines and the proposed approaches

precision the combination is better than any approach and it seems that it has a modestly better F-measure compared to the MFS baseline.

5. Discussion

In the previous section, the results indicate that we can improve over the original automatic generated corpus. We have presented, in addition, several methods which can be used to categorize terms and citations into semantic categories without manually annotated sets. Some semantic groups have been discarded due to problems with the categorizer.

Ambiguous terms like *energy* or *surgery* have profited the most from the filtering. Collocations added to retrieve documents in a reduced number of cases worsened the performance; e.g. *observer* variation. We find several reasons for this: the terms might contain several possible senses not covered in the UMLS Metathesaurus or are simply not valid collocations or have not been properly classified due to mistakes of the semantic group categorizer.

6. Conclusions

In this paper, we have worked on improving the quality of an automatically extracted corpus to train statistical learners to perform WSD. Two approaches have been evaluated. The first one contributed to more specific terms and provided an increase in both precision and recall. The second approach filtered out false positives, generating a large increase in recall. The lack of training data to categorize terms and citations into semantic groups is compensated by corpus statistics.

References

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

Humphrey, S.M., W.J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T.C. Rindflesch. 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*, 57(1):96.

Jimeno-Yepes, A. and A.R. Aronson. 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. http://skr.nlm.nih.gov/papers/references/jimeno_comparison_2010.pdf.

Jimeno-Yepes, A., R. Berlanga-Llavori, and D. Rebholz-Schuhmann. 2009. Comparison of methods for topic template queries in the biomedical domain. In *Languages in Biology and Medicine.*

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

Rosario, B., M.A. Hearst, and C. Fillmore. 2002. 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.

Stevenson, M., Y. Guo, and R. Gaizauskas. 2008. 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.

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

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.