Disambiguating automatically-generated semantic annotations for Life Science open registries

Antonio Jimeno-Yepes[1], Mara Pérez-Catalán[2], and Rafael Berlanga-Llavori[2]

[1] National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
antonio.jimeno@gmail.com
[2] Universitat Jaume I,Castellón, Spain
mcatalan@icc.uji.es,berlanga@lsi.uji.es

Abstract. This paper presents our preliminary evaluation of the automatic semantic annotation of open registries. Conversely to traditional application of semantic annotation to scientific abstracts (e.g., PubMed), open registries contain descriptions that mix terminologies of Computer Science, Biomedicine and Bioinformatics, which makes their automatic annotation more prone to errors. Moreover, the extensive use of acronyms and abbreviations in these registries may also produce wrong annotations. To evaluate the impact of these errors in the quality of the automatically generated annotations we have built a Gold Standard (GS) with single-word annotations. Additionally, we have adapted a knowledge-based disambiguation method to measure the hardness in distinguishing right from wrong annotations. Results show that for some semantic groups the disambiguation can be performed with good precision, but for others the effectiveness is far from being acceptable. Future work will be focused on developing techniques for improving the semantic annotation of these poorly represented semantic groups.

1 Introduction

In recent years, open metadata registries have become a popular tool for researchers trying to locate resources in different domains, mainly in Life Sciences and Open Linked Data. These registries allow users to provide metadata about the resources in order to facilitate their discovery, which can be structured metadata, such as tags or categories, or free text descriptions. Although sophisticated standards have been proposed for annotating the resources, most of the metadata available in the registries are expressed in natural language, which makes more difficult the discovery of these resources in traditional search engines. Descriptions contain useful information about the resources and, moreover, they implicitly describe the features of the resources. Therefore, to facilitate the discovery of the most appropriate web resources, all these metadata has to be normalized in order to be automatically processed.

Semantic annotation techniques are frequently used to normalize the metadata. Semantic annotation (SA) is the process of linking the *entities* mentioned

in a text to their *semantic descriptions*, which are stored in knowledge resources (KRs) such as thesauri and domain ontologies, like UMLS® Metathesaurus® and EDAM ontology [20] in Life Sciences. During the last years, we have witnessed a great interest in massively annotating biomedical information. Most of them are based on dictionary look-up techniques. These approaches try to find in the documents each text span that exactly matches some lexical forms of the terminological resource. Other approaches, like MetaMap [2] and EAGL [22], allow partial matching between text spans and lexical forms. Their main drawback is that precision is usually very low and they suffer from scalability issues. These annotators only base the matching on isolated text spans without taking into account the context of the matching, which is the main source of errors when annotating open collections.

Another issue that has to be taken into account in metadata normalization is that metadata in web resources registries usually contains vocabulary taken from different domains. For instance, in Life Sciences registries, the metadata contains words about medicine, bioinformatics and computers, with a high degree of overlapping between them. However, if the domains are not equally covered by the knowledge resources, some senses of some words can be disregarded and, therefore, the precision of the semantic annotations and, as consequence, also the quality of the retrieved resources may be affected. Thus, the quality of the semantic annotations becomes crucial in the discovery process.

There are two main problems that need to be addressed. One of them is ambiguity, since a term can be mapped to more than one concept or sense. The second one is the lack of coverage of the terminological resources. A term can be ambiguous but this might not be reflected in the terminological resource. As a consequence, there is no guarantee in many cases that even though the mapping is not ambiguous that is correct.

In this paper we study these issues in the context of the semantic annotation of open registries of Life Science resources, using the currently largest biomedical knowledge resource, that is, the NLM's UMLS [5].

2 Methods

We propose to study the effectiveness of unsupervised Word Sense Disambiguation (WSD) approaches. The definition of the concept is turned into a bag-of-words representation in which the words are weighted according to their relevance to the concept and related concepts. This concept profile is compared to the context of the ambiguous word and if it is over a trained threshold according to a similarity measure, then it is assigned the given concept. In this work, the window for the context of the ambiguous word is all the terms in the description of the registry.

The concept profiles are prepared based on the NLM's UMLS [5], which 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.

In the following section, we present the generation of the WSD profiles and present the similarity measures that will be used to compare the concept profiles and the context of the ambiguous words.

2.1 WSD profiles

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 *domain* which could either refer to *works or knowledge without proprietart interest* or, in biology, the *taxonomic subdivision even larger than a kingdom* or *a part of a protein*. The context in which *domain* appears is used to disambiguate it. WSD is an intermediary task which might support other tasks such as: information extraction (IE) [2], information retrieval (IR) and summarization [21].

WSD methods are based either on supervised learning or knowledge-based approaches [23]. 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 (KB) methods rely on models built based on the information available from available knowledge sources. In the biomedical domain, this would include the Unified Medical Language System (UMLS). In this scenario, the candidate senses of the ambiguous word are UMLS concepts. KB methods either build a concept profile [18], develop a graph-based model [1] or rely on the semantic types assigned to each concept for disambiguation [11]. 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.

Due to the scarcity of training data, KB methods are preferred as disambiguation methods. KB methods rely on information available in a terminological resource. Performance of knowledge-based methods depends partly on the knowledge resource, which usually is not built to perform WSD or IR tasks [14].

In our first WSD approach, the context words surrounding the ambiguous word are compared to a profile built from each of the UMLS concepts linked to the ambiguous term being disambiguated. This approach has been previously used by McInnes [18] in the biomedical domain with the NLM WSD corpus.

This algorithm can be seen as a relaxation of Lesk's algorithm [16], which is very expensive since the sense combination might be exponentially large even for a single sentence. Vasilescu et al. [24] have shown that similar or even better performance might be obtained disambiguating each ambiguous word separately.

A concept profile vector has as dimensions the tokens obtained from the concept definition or definitions if available, synonyms, and related concepts excluding siblings.

Stop words are discarded, and Porter stemming is used to normalize the tokens. In addition, the token frequency is normalized based on the inverted *concept* frequency so that terms which are repeated many times within the UMLS will have less relevance.

A context vector for an ambiguous term includes the term frequency; stop words are removed and the Porter stemmer is applied. The word order is lost in the conversion.

2.2 Similarity measures

We have compared the context vector of the term under evaluation (A) and the concept profile vector (B) based on the several similarity measures presented below. The length of the vectors is usually large due to the vocabulary size. But the context and profile vectors only have values for a limited number of entries and the others will have a value of zero.

One of these measures is the cosine similarity, shown in equation 1. The candidate concept with the highest cosine similarity is selected as candidate concept. This approach is used with UMLS based concept profiles [13, 18].

$$Cosine = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

Entailment, presented below, looks at the overlap between the two vectors and normalizes based on the number of tokens in the context vector. Compared to the cosine similarity, the overlap is based on counting the matches between both vectors instead of estimating the dot product. The matches are done considering the non-zero entries. This overlap is normalized by the length of context vector only to avoid a negative impact of a long concept profile.

$$Entailment(A, B) = \frac{|A \cap B|}{|A|} \tag{2}$$

The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. Compared to entailment, the length of the concept profile is considered.

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

Chi-square allows comparing two distributions. In our work, we compare the concept profile to the context vector. Chi-square has been used as a similarity measure in text categorization by Chen et al. [6] and we follow their formulation in this work.

$$\chi^2_v = h \left[\sum_{i=1}^{n} \frac{A_i^2}{sum(A)(A_i + B_i)} + \sum_{i=1}^{n} \frac{B_i^2}{sum(B)(A_i + B_i)} \right] - h \tag{4}$$

$$sum(A) = \sum_{i=1}^{n} A_i \tag{5}$$

$$sum(B) = \sum_{i=1}^{n} B_i \tag{6}$$

$$h = sum(A) + sum(B) \tag{7}$$

2.3 Data set

In this paper, our aim is to analyze the impact of the automatic semantic annotations in the quality of the results of a retrieval system. To do that, we use a dictionary look-up semantic annotator [3] to automatically annotate the metadata of the resources registered in three Life Sciences registries: BioCatalogue [4], myExperiment [10] and SSWAP [9].

The semantic annotator is able to deal with several ontologies in order to cover as much as possible the different vocabularies that appear in the resources descriptions. In this work, the semantic annotator uses as knowledge resources (KRs): UMLS, EDAM (an ontology designed for Life Science open registries), myGrid (reference ontologies of BioCatalogue) and the entries of the Wikipedia that have as category some sub-category of the Bioinformatics category. A detailed description of the semantic annotator can be found in [19].

A preliminary analysis of the automatically generated semantic annotations suggests that concepts matching several words are usually unambiguous and are associated to a right sense. However, single word concepts are much prone to ambiguity and errors.

For this reason, we have manually created a Gold Standard (GS) with those annotations matching a single word. The GS has been curated by two people who have analyzed each combination of concept-word in each semantic annotation in the resources description, selecting the most appropriate concept in each case. The GS contains for each semantic annotation, represented as a triple $(concept, word, contextvector)$, a bit indicating if the sense is correct (1) or not (0). This GS contains 8863 single-word semantic annotations.

The whole catalogue contains 72958 semantic annotations, from which 42686 where annotated only with concepts from UMLS, 12269 were annotated with concepts from UMLS and the other KRs and 18003 were annotated with concepts from the other KRs but not from UMLS.

3 Results

We intend to evaluate the concept profiles and the similarity measures for filtering annotations in our data set. From our data set, we have selected the semantic groups of interest and split the set for each one of the semantic groups sets into 2/3 for training and 1/3 for testing. The semantics groups are the following: CONC (Concepts & Ideas), DISO (Disorders), LIVB (Living Beings) and PHYS (Physiology) as defined in the UMLS Semantic Network [17][3], while the groups CHED (Chemicals & Drugs) and PRGE (Proteins & Genes) follow the definition under the CALBC challenge[4]. CALBC groups definition is closer to our interests compared to the ones defined by the UMLS Semantic Network in these two cases.

Table 1 shows the distribution of semantic annotations of the GS per semantic group. Positive instances are the ones that are labeled with the specified semantic group and the negative ones are instances that should not be labeled with the semantic group. The distribution is usually skewed towards the negative class, i.e. the concept does not represent the correct sense of the word, except for the PRGE group in which the positive examples are more frequent. For example, in the service SMART registered in BioCatalogue, the word *domain* refers to protein domain and it has been annotated with the concepts *C1514562:PRGE*, that refers to the protein domain, and *C1883221:CONC*, that refers to the general concept of domain. Therefore, *C1514562* is the correct concept in this case and it is represented as a positive instance in the GS.

Semantic Group	Training	Positive	Negative	Testing	Positive	Negative
CHED	527	148	379	263	70	193
CONC	2139	598	1541	1068	283	785
DISO	180	6	174	90	4	86
LIVB	408	166	242	203	83	120
PHYS	169	44	125	84	22	62
PRGE	654	460	194	326	232	94

Table 1. Semantic group data set distribution

We would like to be able to decide if an annotation is correct given the measures presented above. We have trained a threshold for each of the measures based on the training set. This threshold is used to decide if the instance should

[3] http://semanticnetwork.nlm.nih.gov/SemGroups
[4] http://www.ebi.ac.uk/Rebholz-srv/CALBC/challenge_guideline.pdf

be labeled with the semantic group or not. The optimization measure has been the F-measure, while other measures could be considered. On the other hand, due to the skewness of the data, other measures as accuracy would not be as effective.

Table 2 shows the filtering performance of the different measures. Overall the similarity measures seem to perform similarly except for chi-square that performs better on average over the other measures. Chi-square shows a larger difference compared to other measures for the LIVB and PHYS semantic groups.

SG	Measure	Threshold	Precision	Recall	F-measure
CHED	chisquare	-3642.0724	0.4898	0.6857	0.5714
	cosine	0.9698	0.5169	0.6571	0.5786
	entailment	0.9269	0.4783	0.6286	0.5432
	jaccard	0.9961	0.4538	0.7714	0.5714
CONC	chisquare	-121.9878	0.2939	0.8693	0.4393
	cosine	1.0000	0.2647	1.0000	0.4186
	entailment	1.0000	0.2647	1.0000	0.4186
	jaccard	1.0000	0.2647	1.0000	0.4186
DISO	chisquare	-41397.6546	0.2500	0.2500	0.2500
	cosine	0.9956	0.1212	1.0000	0.2162
	entailment	0.9407	0.2000	0.5000	0.2857
	jaccard	0.9768	0.0000	0.0000	0.0000
LIVB	chisquare	-3416.2337	0.7349	0.8133	0.7722
	cosine	0.9995	0.4774	0.8916	0.6218
	entailment	0.9302	0.6173	0.6024	0.6098
	jaccard	0.9996	0.4774	0.8916	0.6218
PHYS	chisquare	-884.3186	0.4884	0.9545	0.6462
	cosine	1.0000	0.2619	1.0000	0.4151
	entailment	0.9662	0.3415	0.6364	0.4444
	jaccard	0.9855	0.4063	0.5909	0.4815
PRGE	chisquare	-173.5428	0.7099	0.9914	0.8273
	cosine	1.0000	0.7117	1.0000	0.8315
	entailment	1.0000	0.7117	1.0000	0.8315
	jaccard	1.0000	0.7117	1.0000	0.8315

Table 2. Semantic group results on the test set

4 Discussion

The results are interesting but there is still room for improvement. Among the evaluated measures, chi-square seems to perform better on average compare to the other measures. Cosine has been the preferred similarity measure in many biomedical disambiguation work [13] and would be interesting to evaluate chi-square in similar studies.

The best performing groups are LIVB and PRGE. In the case of LIVB, there are not only the species which have shown already easy to annotate [8], even though this semantic group includes in addition several population groups which seem more difficult to annotate. On the other hand, the best F-measure is obtained when all the cases are annotated as PRGE. This means that in addition to being difficult to annotate, the skewness is in favour of this semantic group.

DISO has a small set of positive cases related to the term *diabetes*. Most of the wrongly assigned terms are abbreviations like CA (California) or SIB (Swiss Bioinformatics Institute). Other mentions like *brain*, have been already identified in previous work [12] and different proposals for lexicon cleansing could be used. This semantic group has a reduced set of annotations which are relevant in our data set, which might indicate that the open registries include almost no mention of diseases.

CONC has the largest number of candidate instances from which only a small part is relevant to this semantic group and appears in large part of the example cases. In this first work, the context vector might be too broad to help decision making over annotations.

PHYS shows a large difference in performance with the chi-square measure. Looking at the examples, there is a limited number of terms used which seem to be always linked to PHYS. Examples of these terms are *pathway*, *transcription* and *transport*. Other terms annotated as PHYS rarely are labeled as PHYS in the gold standard. Among these terms, we find *interactions*, *size* or *status*.

Annotation of chemical entities has already proved to result in low performance [7]. CHED annotations seem to be complicated to filter properly. Again, there are sets of common terms that can be pre-filtered for this domain that in many cases are not related to the topic of interest. Examples of these terms are *products*, *CA* or *date*.

5 Conclusions and Future Work

We have introduced the problem of determining the correct sense of ambiguous terms depending on their context in the semantic annotations in open registries and evaluated the use of knowledge based methods used in disambiguation in the automatic annotation of these registries.

Better performance is required to use the filtered annotations in a retrieval system. We have worked with a large window, all the words in the definition of the registries, in the development of the context vector. A more restrictive window might provide a more focused context. In addition, we have seen that there are terms which seem to have a preferred sense in this data set. Chi-square performs better than other evaluated measures but has not been evaluated in biomedical WSD and could provide better performance than existing work.

We have evaluated knowledge-based WSD methods since, when we started this work, no training data was available. Given the current data set, trained conditional random fields approaches [15] could be evaluated on the annotated set.

Some direct follow-ups of this work are the refinement of particular details of the semantic annotator, such as the detection of locutions as entities that do not have to be annotated, the disambiguation of acronyms, the use of lexical patterns to recognise fragments that are entities as a whole, e.g. the citations, or the disambiguation of single words that are simplifications of multi-words. In addition, we are also considering the use of lexicon cleansing techniques to improve the lexicon.

6 Acknowledgments

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. This work has been also funded by the "Ministerio de Economía y Competitividad", project contract TIN2011-24147.

References

1. Eneko Agirre, Aitor Soroa, and Mark Stevenson. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896, 2010.
2. 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):229–236, 2010.
3. Rafael Berlanga, Victoria Nebot, and Ernesto Jimenez. Semantic annotation of biomedical texts through concept retrieval. In *BioSEPLN 2010*, 2010.
4. Jiten Bhagat, Franck Tanoh, Eric Nzuobontane, Thomas Laurent, Jerzy Orlowski, Marco Roos, Katy Wolstencroft, Sergejs Aleksejevs, Robert Stevens, Steve Pettifer, Rodrigo Lopez, and Carole A Goble. BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic acids research*, 38(Suppl 2):W689–94, jul 2010.
5. O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(Database Issue):D267, 2004.
6. Y.T. Chen and M.C. Chen. Using chi-square statistics to measure similarities for text categorization. *Expert Systems with Applications*, 38(4):3085–3090, 2011.
7. P. Corbett and P. Murray-Rust. High-throughput identification of chemistry in life science texts. *Computational Life Sciences II*, pages 107–118, 2006.
8. M. Gerner, G. Nenadic, and C.M. Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85, 2010.
9. Damian DG Gessler, Gary S Schiltz, Greg D May, Shulamit Avraham, Christopher D Town, David Grant, and Rex T Nelson. SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services. *BMC Bioinformatics*, 10:309, 2009.
10. Carole A. Goble, Jiten Bhagat, Sergejs Aleksejevs, Don Cruickshank, Danius Michaelides, David Newman, Mark Borkum, Sean Bechhofer, Marco Roos, Peter Li, and David De Roure. myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, 38(suppl 2):W677–W682, 2010.

11. S.M. Humphrey, W.J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T.C. 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 (Print)*, 57(1):96, 2006.
12. A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC bioinformatics*, 9(Suppl 3):S3, 2008.
13. A. Jimeno-Yepes and A.R. Aronson. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC bioinformatics*, 11:565, 2010.
14. A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. Ontology refinement for improved information retrieval. *Information Processing & Management*, 2009.
15. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
16. 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.
17. A.T. McCray, A. Burgun, O. Bodenreider, et al. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, (1):216–220, 2001.
18. Bridget 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.
19. M. Pérez-Catalán, R. Berlanga, I. Sanz, and M.J. Aramburu. A semantic approach for the requirement-driven discovery of web resources in the Life Sciences. *Knowledge and Information Systems*, pages 1–20, 2012.
20. Steve Pettifer, Jon Ison, Matus Kalas, Dave Thorne, Philip McDermott, Inge Jonassen, Ali Liaquat, José M. Fernández, Jose M. Rodriguez, INB Partners, David G. Pisano, Christophe Blanchet, Mahmut Uludag, Peter Rice, Edita Bartaseviciute, Kristoffer Rapacki, Maarten Hekkelman, Olivier Sand, Heinz Stockinger, Andrew B. Clegg, Erik Bongcam-Rudloff, Jean Salzemann, Vincent Breton, Teresa K. Attwood, Graham Cameron, and Gert Vriend. The EMBRACE web service collection. *Nucleic Acids Research*, 38(suppl 2):W683–W688, 2010.
21. L. Plaza, A.J. Jimeno-Yepes, A. Díaz, and A. Aronson. Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC bioinformatics*, 12(1):355, 2011.
22. P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(3):658–664, 2006.
23. 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.
24. F. Vasilescu, P. Langlais, and G. Lapalme. Evaluating variants of the Lesk approach for disambiguating words. In *Proceedings of the Conference of Language Resources and Evaluations (LREC 2004)*, pages 633–636, 2004.