Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 247-250 recibido 26-06-10 revisado 03-07-10 aceptado 06-07-10

Semantic annotation of biomedical texts through concept retrieval *

Anotación semántica de textos biomédicos mediante recuperación de conceptos

Rafael Berlanga, Victoria Nebot and Ernesto Jimenez
Departamento de Lenguajes y Sistemas Informaticos
Universitat Jaume I (Spain)
{berlanga,ejimenez,romerom}@lsi.uji.es

Resumen: En este trabajo presentamos una nueva aproximación a la anotación semántica de textos biomédicos basada en la búsqueda de conceptos que mejor cubran el texto que se quiere anotar. La aproximación ha sido evaluada en el contexto de la iniciativa CALBC.
Palabras clave: Recuperación de Conceptos, Anotación Semántica, UMLS®

Abstract: This paper presents a novel approach for the semantic annotation of biomedical texts based on the retrieval of UMLS concepts that best fit with the target text. An evaluation of the approach has been carried out over the CALBC silver standard corpus.
Keywords: Concept Retrieval, Semantic Annotation, UMLS®

1. Introduction

In this paper we introduce the notion of concept retrieval (CR) and how it can be applied to the semantic annotation of biomedical documents. The main idea behind concept retrieval is to regard concepts as documents and the text fragment as queries, so that the problem of semantic annotation is viewed as an information retrieval (IR) task. Thus, the annotation system must first find the most relevant concepts w.r.t. the text words and then select those concepts that best cover the underlying text semantics. The proposed method is therefore fully non-supervised, as it only uses the available lexicon without user intervention.

The notion of concept retrieval has been widely used in the Bioinformatics community, being its main application to categorize documents with different biomedical terminologies, mainly the Medical Subject Headings (MeSH ®) and the Gene Ontology (GO) (Trieschnigg y et al., 2009; Ruch, 2006; Aronson y et al., 2004). These approaches aim to provide a set of relevant concepts as keywords for the target documents, therefore they must be seen as multi-class text classification approaches. In this paper, our goal is to apply CR to the full annotation of biomedical text with very large multi-entity terminologies.

Currently there are few approaches that fully annotates biomedical documents with terminological resources. Most of them, like (Rebholz-Schuhmann y et al., 2008), are based on dictionary look-up techniques. These approaches try to find in the documents each text span that exactly match with some lexical forms of the terminological resource. Although these approaches exhibit good precision numbers, their recall is usually low. Other approaches, like MetaMap (Aronson, 2001) and EAGL (Ruch, 2006), allow partial matching between text spans and lexical forms. The main drawback of these systems is that precision is usually very low. Our ultimate goal is to provide a CR framework in which different ranking models can be implemented and adjusted in order to find a good trade-off between precision and recall adapted to user annotation requirements.

2. Concept Retrieval

Like any other IR model, concept retrieval must measure the similarity between a given query (i.e. a text fragment) and each document of the collection (i.e. concept) in order to give a conceptual cover of the query. Such a measure is usually derived from an IR model (e.g. vectorial, probabilistic, language models, etc.) In our first approach, we have adopted an information-theoretic function which is

* This work has been partially funded by the Spanish National R&D Program (contract TIN2008-01825/TIN)

ISSN 1135-5948

inspired by the matching function defined in (Mottaz et al., 2008) and the word content evidence defined in (Couto, Silva, y Coutinho, 2005). This is defined as follows:

$$sim(C,T) = max_{S \in lex(C)}(ratio(S,T))$$

$$ratio(S,T) = \frac{info(cw(S,T)) - missing(S,T)}{info(S)}$$

$$missing(S,T) = (info(S) - info(cw(S,T)))$$

$info(S)$ measures the relevance of the terms in the string S, and $cw(S,T)$ is the set of terms in common between the concept string S and the text fragment T. It is defined as follows:

$$info(S) = -\sum_{w \in S} log(P(w|UMLS))$$

The function $ratio(S,T)$ defines the ratio between the achieved information evidence for T and the information encoded in the lexicon form S. Finally, $missing(S,T)$ is the amount of information contained in S that have not been covered with T. Notice that conversely to (Mottaz et al., 2008) and (Couto, Silva, y Coutinho, 2005), we take into consideration both covered and uncovered information by T.

The relevance of word is measured by means of its estimated probability within the whole UMLS lexicon (i.e. $P(w|UMLS)$). In this way, highly frequent terms in UMLS contribute little to the final score of the strings containing them. Notice that the final score $sim(C,T)$ is normalized (i.e. it ranges between 0 and 1), and that not all the terms of the lexicon form S must appear in T, but just those that better discriminate the intended concepts. It is worth mentioning that the calculation of this score does not require any parameter except the estimation of $P(w|UMLS)$.

2.1. CR Implementation

The implementation of the concept retrieval system relies on inverted files. Each normalized word has a unique entry which contains the occurrences of the word in each concept string S (hit list). In the hit list, we also store the final score of each concept string (i.e. $info(S)$) in order to speed up the calculation of the similarity function.

Given a text T we first process it to identify the minimum text segments that can potentially allocate a concept. More specifically, we split the text according to some syntactic connectors (e.g. meaningless words like in, while, etc.) and verb forms. Finally, each text segment is treated as a query to be solved with the inverted file. It is worth mentioning that in our first approach we have not used any text chunker based on POS-tagging processing (e.g. OpenNLP, GeniaTagger, etc.), but just a list of splitting words. In the future, we will include part-of-speech tagging to improve the identification of chunk boundaries (e.g. noun phrases).

Once a query Q is given to the system, it returns a list of ranked concepts where each one has also associated a set of words from Q (i.e. matched words).

The final step of our method consists of selecting from this list the concepts that jointly better cover the query. For this purpose, we first group all the concepts having the same score and matched words. Each group is then evaluated according to the following criteria: the ambiguity of the group (i.e. number of different concepts), the maximum gap between the matched words in the text and, the size of the set of matched words. Thus, less ambiguous, more compact (i.e. matched words are closer to each other) and larger matches are preferred candidates for the final result. Finally, a minimum threshold is defined over the score of the retrieved concepts. This threshold allows the algorithm to apply a top-k strategy, which reduces notably the number of concept groups to seek and evaluate. Basically, the top-k strategy consists of estimating a maximum similarity value for the query and then limit the scan of the hit lists up to that value.

2.2. Generating annotations

Given the list of retrieved concepts, the final step consists of identifying the parts of the text associated to each concept and properly annotate them. We adopt the annotation guidelines of the CALBC silver standard corpus (Rebholz-Schuhmann y et al., 2010). Thus, annotations are expressed as XML tags over the original document, without modifying its original contents nor structure. Table 1 shows an example of tagged text using this guideline. Given the list of retrieved concepts $LC(T)$ for the text T, the tagging process is

as follows:

Algorithm 1 Text tagger.

Require: A text T and the ranked retrieved concepts $LC(T)$.
Ensure: The text tagged with a covering of concepts from $LC(T)$.
Initialize $CoveredW = \emptyset$
while $CoveredW \neq T$ and $LC(T) \neq \emptyset$ **do**
pop C_i from $LC(T)$
record the positions of C_i's words in T
append to $CoveredW$ the C_i's matched words
end while
Group concept matches with overlapping positions.
Insert the XML tags for these groups.

3. Experimental Evaluation

CALBC (Rebholz-Schuhmann y et al., 2010) is an initiative to harmonize multiple semantic annotations stemming from different annotators for large biomedical corpora. Unlike other initiatives like BioCreative and Genia, CALBC is aimed at providing a silver standard corpus (SSC) useful for text-mining tasks. The aim of a SSC is similar to golden standards (GS) in the sense that it can be taken as a reference of quality for the annotators, but it also provides a means of comparing the output of disparate annotation methods and integrate them into a new SSC. We must point out that CALBC corpus is currently aimed to *named entity recognition*, that is, it identifies text spans where entities of semantic groups occur. However, our approach also performs *named entity resolution*, that is, it assigns a concept identifier to each text span. As a consequence, the current evaluation only focus on the named entity recognition task. Nevertheless, as one of the main aspects we wanted to explore was scalability, we considered this evaluation very interesting.

The number of entities retrieved by our approach (CR stands for Concept Retrieval) is compared with SSC I in Table 2. The semantic groups that where included in the SSC I are: chemical products (CHED), protein-gene (PRGE), disorders and diseases (DISO) and species (SPE). As it can be noticed, the number of retrieved entities is always higher than that of SSC I for all the semantic groups, except for SPE.

Semantic Group	SSC I	CR
CHED	228,622	602,317
PRGE	275,235	531,729
DISO	300,637	332,413
SPE	317,211	310,591

Cuadro 2: Number of entity mentions identified in the corpus.

Sem. Group	P (exact)	R (exact)	F1
CHED	10.0	26.3	14.5
PRGE	9.4	18.1	12.4
DISO	28.5	31.5	29.9
SPE	39.2	38.4	38.8

Cuadro 3: Results for the exact match between CR and SSC annotations.

In order to check the overlap of the annotations obtained in both SSC I and our approach, we show the recall/precision results reported by CALBC organization. In CALBC there are two matching approaches for comparing the annotations of two systems: exact and approximate measures. With the former one, annotation agreement is achieved when the strings of the annotation are exactly the same[1]. With the latter one, annotated strings are compared with the cosine measure by weighting the involved words with their IDF estimated from a large collection. The threshold used to decide the agreement is 0.8, hence it is named `cos98`.

Tables 3 and 4 show the results obtained for these two measures. We can conclude from them that both sets of annotations are quite different from each other, being their overlap quite low. As both precision and recall values increase when the approximate comparison method is applied, we can also conclude that a good percentage of the disagreement

[1]After removing stop words.

Sem. Group	P (app)	R (app)	F1
CHED	11.5	30.3	16.7
PRGE	13.4	25.9	17.6
DISO	34.1	30.8	32.4
SPE	42.7	41.8	42.2

Cuadro 4: Results for the approximate match (cos98) between CR and SSC annotations.

```
<e id="UMLS:C1709323:T062::1,2"><w id="1">Open</w> <w id="2">label</w></e>
<e id="UMLS:C0282460:T062::1,2,3"><w id="1">phase</w> <w id="2">II</w><w id="3">trial</w></e>
of <e id="UMLS:C0205171:T081">single</e>, <e id="UMLS:C0205385:T080">ascending</e>
<e id="UMLS:C0439568:T079">doses</e> of MRA in
<e id="UMLS:C0007457:T098|UMLS:C0043157:T098">Caucasian</e><e id="UMLS:C0008059:T100">children</e>
with <e id="UMLS:C0205082:T080">severe</e>
<e id="UMLS:C1384600:T047::1,2,3,4|UMLS:C0682057:T100::2"><w id="1">systemic</w> <w id="2">juvenile</w>
<w id="3">idiopathic</w><w id="4">arthritis</w></e>: proof of principle of the
<e id="UMLS:C1707887:T062">efficacy</e> of
<e id="UMLS:C0063717:T116,T129,T192::1,2"><w id="1">IL-6</w> <w id="2">receptor</w> </e>
<e id="UMLS:C0332206:T169">blockade</e> in this
<e id="UMLS:C0332307:T080|UMLS:C0455704:T170">type</e> of arthritis and demonstration of
<e id="UMLS:C0439590:T079">prolonged</e> <e id="UMLS:C0205210:T080">clinical</e> improvement.</s>
```

Cuadro 1: Example of tagged text with CALBC-like annotations.

is due to the boundaries of the annotations. However, much disagreement probably stems from the number of recognized entities in the text (Table 2).

Finally, with respect to the partial results per semantic group, the disagreement is higher for chemical and protein-gene groups, whereas it is lower for species. This result was already reported in the CALBC workshop for all the participants, so it indicates that there is much more variation in the entity representations of CHED and PRGE groups than in DISO and SPE.

4. Conclusions

In this work we have shown how concept retrieval can be used to annotate large corpora of biomedical texts, although its former motivation was annotating semi-structured data from medical protocols. The evaluation over the SSC of the CALBC initiative has shown that CR can contribute with new kinds of annotations to the SSC.

In the future work we plan to evaluate CR over golden standards in order to get a more precise measure of the quality of the annotations. First experiments with the GS presented in (Mottaz et al., 2008) shows an agreement around 62.0 % for the entities tagged in the *disease* field of 100 Uniprot protein database entries, which can be considered a good percentage. We expect to present in the workshop more results over other existing gold standards.

As future work, we plan to perform exhaustive experiments to measure the impact of each component of the annotation system, mainly in the scoring function and the text segmentation strategy.

Acknowledgments. We would like to thank the anonymous reviewers for their helpful comments.

Bibliography

Aronson, Alan R. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. En *Proccedings of the 2001 AMIA Symposium*, páginas 17–21.

Aronson, Alan R. y et al. 2004. The NLM indexing initiative's medical text indexer. En *In Proceedings of the 11th World Congress on Medical Informatics Demner-Fushman and Lin Answering Clinical Questions (MEDINFO 2004*, páginas 268–272.

Couto, Francisco M., Mário J. Silva, y Pedro Coutinho. 2005. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(S-1).

Mottaz, Anaïs, Yum Lina Yip, Patrick Ruch, y Anne-Lise Veuthey. 2008. Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics*, 9(S-5).

Rebholz-Schuhmann, Dietrich y et al. 2008. Text processing through web services: calling whatizit. *Bioinformatics*, 24(2):296–298.

Rebholz-Schuhmann, Dietrich y et al. 2010. CALBC silver standard corpus. *J Bioinform Comput Biol*, 8(1):163–79.

Ruch, Patrick. 2006. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658–664.

Trieschnigg, Dolf y et al. 2009. Mesh up: effective mesh text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418.