An evaluation of the concept retrieval annotation for Spanish-English CLEFER parallel corpora

Rafael Berlanga[1], Antonio Jimeno-Yepes[2], María Pérez-Catalán[1], and Dietrich Rebholz-Shuchmann[3]

[1] Department of Languages and Computer Systems
Universitat Jaume I, Castelló, Spain
email:{berlanga,maria.perez}@uji.es
[2] National ICT Australia
email: antonio.jimeno@gmail.com
[3] Department of Computational Linguistics,
University of Zürich, Ch
rebholz@ifi.uzh.ch

1 Introduction

This paper presents a study about the use of the *concept retrieval* annotation method for parallel corpora. The concept retrieval annotation method (CRA) consists of considering concepts as documents and text chunks as queries [1]. Concepts with higher similarity to text chunks are considered for generating the final semantic annotation. CRA makes use of an existing knowledge resource (KR) from which lexicons are extracted to perform the semantic annotation. Until now, CRA has been applied to mono-lingual scenarios showing a good performance over both very large collections (e.g., CALBCII-SSC[4]) and very large lexicons (e.g., UMLS® [2]). We have also applied this semantic annotator to different tasks in Biomedicine such as resource discovery [3], relation extraction [4], and sicentific bibliography analysis [5].

In this work, we will apply CRA in a bi-lingual scenario. For this purpose, we make use of the provided lexicons at CLEFER workshop. More specifically, we have made use of the English and Spanish lexicons. In this extended abstract, we first summarize the main features of CRM as a cross-lingual annotator, and then obtained results over the two provided parallel corpora, EMEA and MEDLINE®.

2 Concept retrieval-based semantic annotation

Performing the semantic annotation of a document D consists of finding mappings between text chunks of D (i.e., sequences of adjacent terms), and the concepts provided by a knowledge resource (KR) that best semantically describes the contents of D. As concepts of a KR are usually expressed as noun phrases, text chunks are usually associated to these syntactic structures. We assume that

[4] http://www.ebi.ac.uk/Rebholz-srv/CALBC/

there exists a function $lex_{KR}(C)$ that returns the set of strings describing the concept C. This set of strings can contain different lexical variants of C, synonyms of these variants, and a short definition of the concept.

Given a text chunk $T = (w_1 \cdots w_n)$, and the concept C of the KR, the retrieval score of C w.r.t. to T is calculated as follows [1]:

$$sim(T, C) = max_{s \in lex(C)} \frac{info(s \cap T) - info(T - s)}{info(s)}$$

The function $info(s)$ provides the information the string s brings, which is calculated as $info(s) = \sum_{w \in s} -log(p(w|Background))$.

The retrieval of candidate concepts is efficiently performed by using an inverted file where each entry is a vocabulary word, and the hit list contains the concept strings containing the word. In this way, the text chunk T is executed as a query over this inverted file, and the retrieved concept strings are ranked according to $sim(T, C)$. Finally, the top-ranked concepts that best cover the words in T are included in the semantic annotation of T.

We propose a simple strategy to annotate texts given a multi-lingual KR:

- Build a different inverted file for each language supported by the KR.
- Define a series of simple lexical rules to generate variants from one language to the other (e.g., proteína $\longrightarrow$ protein).
- Fetch the query to each inverted file with the variants corresponding to its language.
- Return the set of all concepts retrieved by each lexicon.

In the multi-lingual scenario we also have to estimate the word probabilities in large text collections for each language. Fortunately, there exist several publicly available resources providing such word estimations [5]. We have performed Word Sense Disambiguation (WSD) based on the MRD (Machine Readable Dictionary) method [6] built on the UMLS2012AB, for both English and Spanish. In EMEA, the context for disambiguaton is the *document* instead of the *unit*, since broader context has shown to produce better disambiguation results. In the MEDLINE annotation, there is only one unit per document.

3 Results

Table 1 shows the main features of the annotated collections. Annotated collections provided at CLEFER are indicated with SSC (Silver Standard Corpus), and they are in English. Annotations are calculated as the number of text chunks having associated some concept. The average size of an annotation is the average of the number of words of annotated text chunks. We also measure the percentage of ambiguous annotations, which are those having more than one entity type associated to the text chunk. In general, English collections generate more ambiguous annotations than the Spanish ones. However, this result is

[5] http://invokeit.wordpress.com/frequency-word-lists/

Annotated Collection	Documents/Units	Annotations	Ann. Avg. size	Ambiguity
EMEA SSC	879/364005	971715	1.25	5.2%
EMEA EN	879/364005	427013	1.3	10.8%
EMEA EN (Ed)	879/364005	373971	1.3	5.2%
EMEA ES (Ed)	895/140552	433671	1.5	5.6%
MEDLINE SSC	1593546	4101813	1.3	5.2%
MEDLINE EN (Ed)	1593546	3529800	1.5	5.6%
MEDLINE ES (Ed)	247655	610636	1.5	7.0%

Table 1. Features of the annotations generated for the selected datasets.

mainly due to the higher noise of the English lexicon. We noticed that many ambiguous annotations were derived from contextual descriptions. To alleviate this problem, we developed a simple heuristic to detect these false ambiguities, and accordingly edited the lexicons. Thus, collections EMEA EN/ES (Ed) and MEDLINE EN/ES (Ed) have been annotated using the edited lexicons. Notice that for the EMEA English collection, this heuristic has reduced notably the ambiguity degree of the annotations (compare the second and third rows of the table).

Table 2 shows the concept overlap at both collection and the aligned unit levels. As it can be noticed, the edition of the lexicon increases the overlap between the collections. This is because wrongly annotated concepts are unlikely to appear in the parallel collection. It must be also noticed that overlap at collection level is much higher than at unit level.

Collection	Collection level	Unit level
EMEA EN/SS	80%	75%
EMEA ES/SS	80%	58%
EMEA EN/ES	79.7%	53%
MEDLINE EN/SS	76.9%	68%
MEDLINE ES/SS	46.7%	56%
MEDLINE EN/ES	42.0%	51%

Table 2. Overlap of concepts at collection and aligned unit levels (% common CUIs).

Acknowledgements

This work has been partially funded by the Spanish National R&D Programme project with contract number TIN2011-24147 of the “Ministerio de Economía y Competitividad”, and the EU STREP project grant 296410 (“Mantra”) under the 7th EU Framework Programme within Theme “Information Content Technologies, Technologies for Digital Content and Languages” [FP7-ICT-2011-4.1].

It has been also supported by Australian Federal and Victoria State Governments and the Australian Research Council through the ICT Centre of Excellence program, National ICT Australia (NICTA).

References

1. Berlanga, R., Nebot, V., Jiménez, E.: Semantic annotation of biomedical texts through concept retrieval. Procesamiento del Lenguaje Natural **45** (2010) 247–250
2. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research **32**(Database Issue) (2004) D267
3. Pérez, M., Berlanga, R., Sanz, I., Aramburu, M.J.: A semantic approach for the requirement-driven discovery of web resources in the life sciences. Knowl. Inf. Syst. **34**(3) (2013) 671–690
4. Nebot, V., Berlanga, R.: Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. Knowledge and Information Systems (2012) 1–25
5. Berlanga, R., Jiménez-Ruiz, E., Nebot, V.: Exploring and linking biomedical resources through multidimensional semantic spaces. BMC Bioinformatics **13**(S-1) (2012) S6
6. Jimeno-Yepes, A., Aronson, A.: Knowledge-based biomedical word sense disambiguation: comparison of approaches. BMC bioinformatics **11:565** (2010)