Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 251-254 recibido 26-06-10 revisado 03-07-10 aceptado 06-07-10

Using Wikipedia for term extraction in the biomedical domain: first experiences

Utilización de Wikipedia para la extracción de términos en el dominio biomédico: primeras experiencias

Jorge Vivaldi
Applied Linguistic Institute, UPF
Roc Boronat 138, Barcelona, Spain
jorge.vivaldi@upf.edu

Horacio Rodríguez
Software Department, UPC
Jordi Girona Salgado 1-3, Barcelona, Spain
horacio@lsi.upc.edu

Resumen: Presentamos un sistema de extracción de términos que usa la Wikipedia como fuente de información semántica. El sistema ha sido probado en un corpus médico en español. Comparamos los resultados usando un módulo de un extractor de términos híbrido y un módulo equivalente que utiliza la Wikipedia. Los resultados demuestran que este recurso puede utilizarse para esta tarea.
Palabras clave: extracción de términos, ontologías, Wikipedia.

Abstract: We present a term extractor that uses Wikipedia as an semantic information source. The system has been tested on a Spanish medical corpus. We compare the results obtained using a module of a hybrid term extractor and an equivalent module that use the Wikipedia. The results show that this resource may be used for this task.
Keywords: term extraction, ontologies, Wikipedia.

1 Introduction and Motivation

Terms are usually defined as lexical units that designate concepts in a thematically restricted domain. A main problem concerning terms regards their detection. This is a difficult task because, in a given language, "terms and words adopt the same word formation rules".

Term extraction (TE) can be seen as semantic annotation task because it provides machine-readable information based on meaning. Ways to attack the problem depend on the available resources for each language. Some languages have large ontologies and/or term repositories that can be used for reference while other languages have to rely on other procedures. For the former, the procedure starts by parsing the text into noun phrases and then tries to map it to concepts of the domain. For systems lacking these resources, typical approaches involve linguistic/statistical strategies with results not fully satisfactory (Cabré et al., 2001). One of the reasons of such behaviour is the lack semantic knowledge. Notable exceptions are TRUCKS (Maynard, 1999) and YATE (Vivaldi, 2001) that use UMLS[1] and EuroWordNet (EWN)[2] respectively. For medical term extraction, we have to quote FASTR (Jacquemin, 2001), and Metamap (Aronson and Lang, 2010).

EWN is a general-purpose multilingual lexical database; so, we need to determine which areas belong to the domain of interest. It may be done by defining a set of domain markers (DM), i.e, EWN nodes whose attached strings belong to the medical domain, as well as the variants of all (or at least most of) its hyponyms. Initially, DMs were selected manually starting with a set of seed words for the domain, looking for the corresponding nodes in EWN and exploring their environment.

As this procedure is costly and difficult to scale up, (Vivaldi and Rodríguez, 2004) faced the problem using a glossary of the domain. Also (Vivaldi and Rodríguez, 2010) explored the possibility of using Wikipedia[3] (WP) as a KS because of its wide coverage of domain vocabulary for some language. As these results were encouraging, we decided to apply such methodology in the medical domain.

[1] http://www.nlm.nih.gov/research/umls/
[2] http://www.illc.uva.nl/EuroWordNet/
[3] http://www.wikipedia.org/

ISSN 1135-5948

2 Methodology

The basic idea of our approach is: given a document and the corresponding set of TC, to compare the results obtained either 1) using Domain Coefficient (DC) and a set of Domain Markers (DM) as defined by YATE (therefore using EWN) with 2) a similar approach using WP (see below) instead of EWN. For this experiment we used a single DM that corresponds to the category of WP that coincides with the domain name (Medicine). The whole methodology is shown in Figure 1.

The key is to explore WP in order to calculate a DC equivalent to those obtained using EWN. For a given TC, the basic procedure consists of i) finding a WP page that corresponds to such TC, ii) finding all WP categories associated to such page and iii) exploring WP following recursively all super categories links found in the previous step to the reach the domain border.

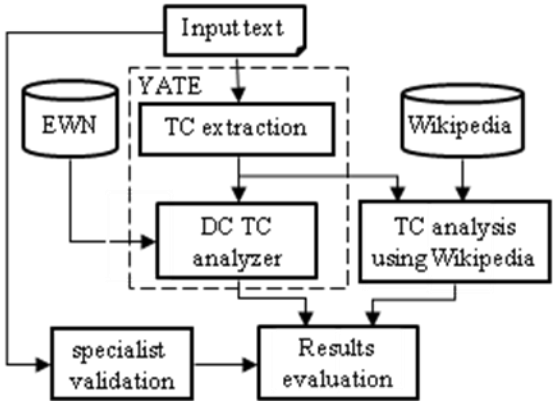

Figure 1. General overview

Using the information collected during this exploration we defined several ways to calculate the DC for a given term *t*.:

1. DC based on the number of path. This coefficient is defined as follows:

$$CDnc(t) = \frac{NP_{domain}(t)}{NP_{total}(t)} \quad (1)$$

where $NP_{domain}(t)$ number of paths to the domain category
$NP_{total}(t)$ number of paths to the top

2. DC based in the number of single steps. This coefficient is defined as follows:

$$CDlc(t) = \frac{NS_{domain}(t)}{NS_{total}(t)} \quad (2)$$

where $NS_{domain}(t)$ number of steps to the domain category
$NS_{total}(t)$ number of steps to the top

3. DC based on the average length paths. This coefficient is defined as follows:

$$CDlmc(t) = \frac{ALP_{domain}(t)}{AVP_{total}(t)} \quad (3)$$

where $NP_{domain}(t)$ average path length to the domain category
$NP_{total}(t)$ average path length to the top

Figure 2 show a simplified sample of the WP organization around the Spanish term *sangre* (blood). The domain category chosen as DB is Medicine (shaded oval).

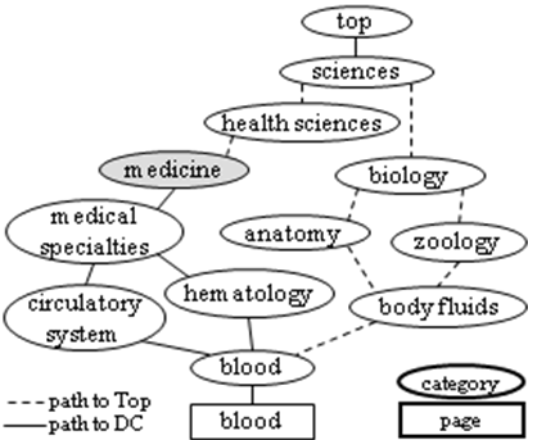

Figure 2. Sample of Spanish WP category tree for the term "blood"

For the last three DC methods an additional step consisting of building a set of WP categories belonging to the domain (*CatDomSet*) is needed. For doing so, we start at the top domain category and traverse top down the category graph, avoiding cycles, collecting all subcategories. From this set we remove all proper names and service classes.

For cleaning the set we measure the medicalhood of both categories and pages belonging to such categories and use thresholds for removing undesirable categories (Vivaldi, Rodriguez, 2010). In our case an initial set of 2431 categories was reduced to 839.

Once *CatDomSet* is built, the last three DC methods can be applied. For each TC, *t*, occuring in WP we obtain its page P_t (performing a disambiguation process when needed). Then we get the set of categories P_t belongs to. We split this set into three subsets: the categories belonging to *CatDomSet*, the categories not belonging to *CatDomSet* and the

categories we name "Neutral Categories", i.e, categories added to WP by convenience for structuring the database or due to its encyclopaedic character (e.g. "scientists by country", …) or categories used temporally for monitoring the state of the page (e.g. "Articles to be split",...). Neutral categories are simply not taken into account for counting.

$P_t Score$ is defined as the ratio between the number of categories belonging to *CatDomSet* and the total number of categories excluding neutral ones. $inP_t Score$ and $outP_t Score$ consider the sets of pages pointing to P_t (for $inP_t Score$) and pointed from P_t (for $outP_t Score$). All these pages are scored in the same way of $P_t Score$. Then $inP_t Score$ and $outP_t Score$ are computed as average of the corresponding scores of pages belonging to the corresponding sets.

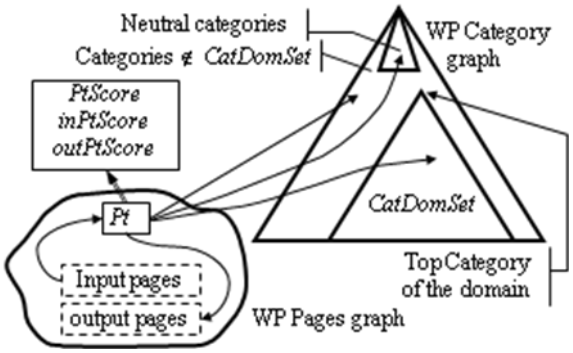

Figure 3. WP additional filtering

For combining the results of these methods we have learned a decision tree classifier using Weka (Witten & Eibe, 2005)[4]. We have used as features the 6 DC methods defined above, the syntactic class of *t* and the type of P_t.

3 Results

We tested the behaviour of the DCs defined in Section 2 using a subset of the IULA's LSP Corpus (100 Kwords)[5]. This document has been linguistically processed as usual in most of the NLP tasks and we evaluate the results using the standard measures of precision and recall.

For evaluation we perform two set of tests: i) we evaluate the behaviour of the DC -as defined in (1), (2) and (3)- and ii) we evaluate the behaviour of the system using the additional information obtained from WP. In both cases we evaluate all patterns[6] together but in the former we evaluate also each pattern individually. The results obtained using just the DCs obtained from WP are shown in Figures 3-5 for the main patterns individually while Figure 7 shows result for all patterns together.

As can be seen above, for all patterns the results obtained using YATE is slightly better. Such behaviour is due to the EWN version used by the TE was adapted to the domain. But the difference is not too high as may be expected. We analyze the results for each pattern and the results may be summarized as follows:

- Pattern N: The difference between EWN and WP varies from 10% (*CDnc*) to 25% (*CDlmc*). In spite of this we point that *CDnc* ranks very well TC not present in EWN.
- Pattern NJ: The behaviour of CDs is similar and differences are around 25%. TC like *historia clínica* (medical record), or *signo clínico* (medical sign) are classified better that by using EWN. Some terms are detected by WP but not by EWN and viceversa.
- Pattern NPN: In this case the performance of all WP based CDs is better than those using EWN. The reason is that YATE performance is very poor for this pattern due to EWN peculiarities. Besides, WP contains many terminological units like *grupo de riesgo* (risk group) and *índice de mortalidad* (mortality rate) that get the maximum value with *CDnc* but very low values using EWN. Only a few terms are included in WP. From 910 candidates only 14 have a positive CD and 39 candidates occur in WP.
- All patterns: considering this global performance the difference in precision among EWN and any of the WP-based CDs is below 5% for a 30% of recall.

As usual, the list of terms manually tagged is troublesome due to completeness and criteria differences. It leaves aside some correct terms as *epitelio* (epithelium) or *medicina interna* (internal medicine).

Figure 8 presents the results of the combination. The basic classifier learned consisted of 20 rules. We scored each rule with its individual accuracy on the set of 4000 TCs given by WEKA. The rules were then sorted by decreasing accuracy and all the subsets of more

[4] http://www.cs.waikato.ac.nz/ml/weka/

[5] Manual annotation resulted on 1444 terms from 5251 candidates.

6 Terms are built mostly using the following linguistics patterns: noun, noun-adjective and noun-preposition-noun.

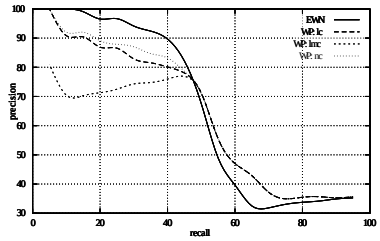

Figure 4. Noun pattern

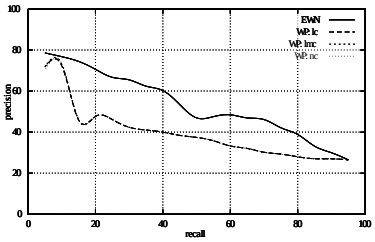

Figure 5. Noun-Adjective pattern

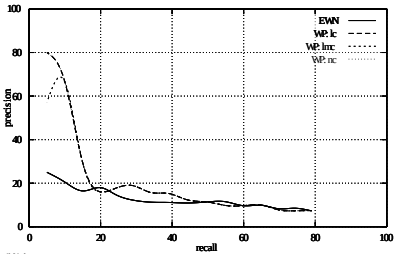

Figure 6. Noun-Prep-Noun pattern

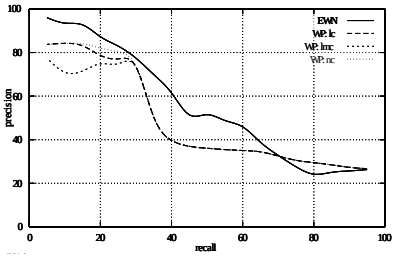

Figure 7. All patterns

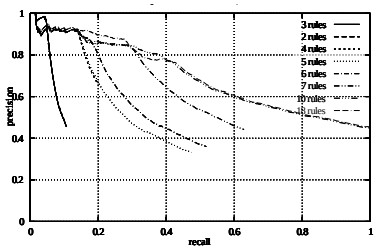

Figure 8. Results using decision trees

accurate rules from 1 to 20 rules were applied with the results shown in Figure 8. The combination consistently outperforms the 3 WP based classifiers, except for the low coverage zone of the EWN based classifier.

4 Conclusions

The methodology proposed in this paper opens the possibility to do TE on biomedical texts using WP, a well known resource available not only for English but also for other languages. Although WP is not a domain specific resource, the results obtained are pretty good. As a matter of fact, the expected results fully depend of the quality and completeness of WP (and other NLP specific resources) in a given language.

In the future we plan to apply this methodology to other languages as well as to improve the integration of WP in a TE system. Also we plan to improve the exploration of the WP Category tree using Bayesian networks.

Bibliografía

Aronson A. & F. Lang, 2010. An overview of MetaMap: historical perspective and recent advances. JAMIA 2010 17:229-236.

Jacquemin C., 2001. *Spotting and discovering terms through natural language processing*. MIT Press.

Maynard, D., 1999. *Term recognition using combined knowledge sources*. PhD Thesis. Manchester Metropolitan University

Vivaldi and Rodríguez, 2010, Finding Domain Terms using Wikipedia. In *Proceedings of the 7th LREC International Conference,*

Vivaldi and Rodríguez, 2004. Automatically selecting domain markers for terminology extraction. In *Proceedings of the 4th LREC International Conference.*

Vivaldi, J., 2001. Extracción de Candidatos a Término mediante combinación de estrategias heterogéneas. PhD thesis. UPC.

Witten, I.H. and F., Eibe, 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann