Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 129-136 recibido 14-04-14 revisado 15-07-14 aceptado 15-07-14

Boosting Terminology Extraction through Crosslingual Resources

Mejora de la extracción de terminología usando recursos translingües

Sergio Cajal, Horacio Rodríguez
Universitat Politècnica de Catalunya, TALP Research Center
Jordi Girona Salgado 1-3 edifici Omega D-316
Campus Nord 08034 Barcelona (Spain)
scajal@gmail.com, horacio@lsi.upc.edu

Resumen: La extracción de terminología es una tarea de procesamiento de la lengua sumamente importante y aplicable en numerosas áreas. La tarea se ha abordado desde múltiples perspectivas y utilizando técnicas diversas. También se han propuesto sistemas independientes de la lengua y del dominio. La contribución de este artículo se centra en las mejoras que los sistemas de extracción de terminología pueden lograr utilizando recursos translingües, y concretamente la Wikipedia y en el uso de una variante de PageRank para valorar los candidatos a término.
Palabras clave: Extracción de terminología. Procesamiento translingüe de la lengua. Wikipedia, PageRank

Abstract: Terminology Extraction is an important Natural Language Processing task with multiple applications in many areas. The task has been approached from different points of view using different techniques. Language and domain independent systems have been proposed as well. Our contribution in this paper focuses on the improvements on Terminology Extraction using crosslingual resources and specifically the Wikipedia and on the use of a variant of PageRank for scoring the candidate terms.
Keywords: Terminology Extraction, Wikipedia, crosslingual NLP, PageRank

1 Introduction

Terminology Extraction is an important Natural Language Processing, *NLP*, task with multiple applications in many areas. Domain terms are a useful mean for tuning both resources and *NLP* processors to domain specific tasks. The task is important and useful but it is also challenging. In (Krauthammer, Nenadic, 2004), it has been said that "terms identification has been recognized as the current bottleneck in text mining and therefore an important research topic in *NLP*".

Terms are usually defined as lexical units that designate concepts in a restricted domain. Term extraction (or detection) is difficult because there is no formal difference between a term and a non terminological unit of the language. Furthermore, the frontier between terminological and general units is not always clear and the belonging to a domain is more a fuzzy than a rigid function. (Hartmann, Szarvas and Gurevych, 2012) present the lexical units in a two dimensional space where *x* axe refers to *domainhood*, represented as a continuous, and *y* axe to *constituency* of the linguistic unit, i.e. single words and multiword expressions, *MWE*, (2-grams, 3-grams, etc.). Several types of *MWE* can be considered such as idioms, "kick the bucket", particle verbs, "fall off", collocations, "shake hands", Named Entities, "Los Angeles", compound nouns, "car park", some of which are compositional and other not. Obviously not all the *MWE* are terminological and not all the terms are *MWE*[1].

In this paper we prefer to refer to terms as term candidates (*TC*). As pointed out above, *TC*

[1] Many authors claim that most terms are MWE-From our experience we think that almost half of the TC extracted are single words.

ISSN 1135-5948

can be atomic lexical units or *MWE* composed by atomic units (usually named basic components of the term). There are some properties that must hold for a given *TC* in order to be considered a term: i) *unithood*, ii) *termhood* and iii) specialized usage. *Unithood* refers to the internal coherence of a unit: Only some sequences of POS tags can produce a valid term, N (e.g. "Hepatology" in the Medical domain), NN (e.g. "Blood test"), JN (e.g. "Nicotinic antagonist"), etc. and these combinations are highly language dependent), *termhood* to the degree a *TC* is related to a domain-specific concept and specialized usage (general language versus specialized domain). It is clear that measuring such properties is not an easy task. They can only be measured indirectly by means of other properties easier to define and measure like frequency (of the *TC* itself, its basic components or in relation to general domain corpus), association measures, syntactic context exploration, highlighting and/or structural properties, position in an ontology, etc.

We present in this paper a term ranker aimed to extract a list of *TC* sorted by *termhood*. Our claim is that the system is language and domain independent. In fact nothing in our approach depends on the language or the domain. The experiments and evaluation are carried out in two domains, *medicine* and *finance* and four languages: English, Spanish, Catalan, and Arabic.

Our approach is based on extracting for each domain the *TC* corresponding to all the languages simultaneously, in a way that the terms extracted for a language can reinforce the corresponding to the other languages. As unique knowledge sources we use the wikipedias of the involved languages.

Following this introduction, the paper is organized as follows. In section 2 we describe some recent work done in this area. Section 3 describes the methodology that we use to obtain new terms while section 4 describes the experiments carried out as well as its evaluation. Finally, in section 5, we present some conclusions and directions for future work.

2 Related work

Term extraction, *TE*, and related tasks (Term ranking, Named Entity Recognition, *MWE* extraction, lexicon and ontology building, multilingual lexical extraction, etc.) have been approached typically using linguistic knowledge, as in (Heidet al, 1996), or statistical strategies, such as ANA (Enguehard, Pantera, 1994), with results not fully satisfactory, see (Cabré, Estopà, Vivaldi, 2001) and (Pazienza. Pennacchiotti, Zanzotto, 2005). Also, *TE* systems often favor recall over precision resulting in a large number of *TC* that have to be manually checked and cleaned.

Some approaches combine both linguistic knowledge and Statistics, such as TermoStat (Drouin, 2003), or (Frantzi, Ananiadou and Tsujii, 2009), obtaining clear improvement. A common limitation of most extractors is that they do not use semantic knowledge, therefore their accuracy is limited. Notable exceptions are Metamap (Aronson, Lang, 2010) and YATE (Vivaldi, 2001).

Wikipedia[2], *WP*, is by far the largest encyclopedia in existence with more than 32 million articles contributed by thousands of volunteers. *WP* experiments an explosive growing. There are versions of *WP* in more than 300 languages although the coverage (number of articles and average size of each article) is very irregular. For the languages covered by the experiments reported here the size of the corresponding *WPs* are 4,481,977 pages in English, 1,091,299 in Spanish, 425,012 in Catalan, and 269,331 in Arabic. A lot of work has been performed for using this resource in a variety of ways. See (Medelyan et al, 2009) and (Gabrilovich, Markovitch, 2009) for excellent surveys.

WP has been, from the very beginning, an excellent source of terminological information. (Hartmann, Szarvas and Gurevych, 2012) present a good survey of main approaches, see also (Sabbah, Abuzir, 2005). Both the structure of *WP* articles (infoboxes, categories, redirect pages, input, output, and interlingual links, disambiguation pages, etc.) and their content have been used for *TE*. Figure 1 presents the bi-graph structure of *WP*. This bi-graph structure is far to be safe. Not always the category links denote belonging of the article to the category; the link can be used to many other purposes. The same problem occurs in the case of links between categories, not always these links denote hyperonymy/hyponymy and so the structure shown in the left of figure 1 is not a real taxonomy. Even worse is the case of inter-

[2] https://www.wikipedia.org/

page links where the semantics of the link is absolutely unknown.

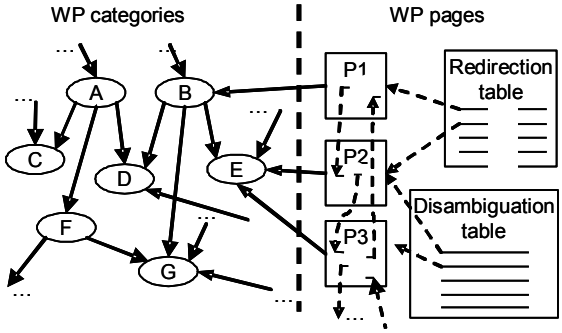

Figure 1: The graph structure of Wikipedia

Nakayama and colleagues, (Erdmann et al, 2008), and (Erdmann et al, 2009) face the problem of bilingual terminology extraction mainly using the interlingual links of *WP*, while (Sadat, 2011) uses, as well, the context of words and the Wikitionary[3]. Gurevych and collegues, (Wolf, Gurevych, 2010), (Niemann, Gurevych, 2011, map *WP* and WordNet[4], *WN*. Vivaldi and Rodríguez propose in (Vivaldi, Rodríguez, 2011) to use *WP* for extracting and evaluating term candidates in the medical domain, and in (Vivaldi, Rodríguez, 2012) propose to obtain lists of terms a multilingual/multidomain setting. (Alkhalifa, Rodríguez, 2010) use *WP* for enriching the Arabic WordNet with *NE*. The approaches more related to ours are those of Vivaldi and Rodríguez, both use WP categories and pages and their relations as knowledge sources, but there are clear differences: our use of interlingual links and the way of scoring candidates by means of the modified PageRank algorithm.

3 Our approach

The global architecture of our approach is displayed in Figure 2. As we can see it consists of 6 steps that are applied for each of the domains as detailed below. Let *d* be the domain considered (as we will see in Section 4 our experiments and evaluation have been carried out for medicine and finance). We will note WP^l the wikipedia for language *l* (*l* ranging on the four languages considered, i.e. en, sp, ca, ar).

[3] http://www.wiktionary.org/

[4] http://wordnet.princeton.edu/

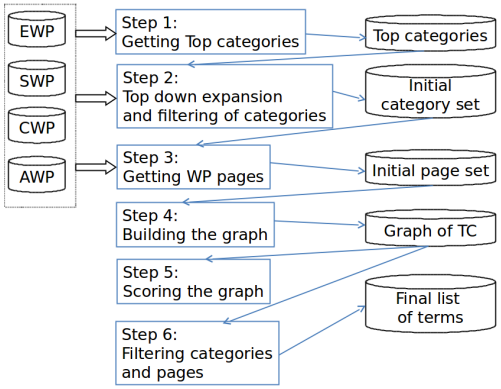

Figure 2: Architecture of our approach

As a preparatory step we downloaded the required *WP*s from the *WP* dumps site[5] and then we used the *JWLP*[6] toolbox, (Zesch, Müller, Gurevych, 2008) for obtaining a *MYSQL* representation of the *WP*s and the interlingual links. We then looked for the Top of the *WP* category graph (*topCat*), that for English *WP* corresponds to "Articles"[7]. Further on we enriched the *WP* category graph with the depth respect to *topCat* of all the categories. We have also downloaded the tables corresponding to the interlingual links. Although these links present problems of lack of reciprocity and inconsistency, see (De Melo, Weikum, 2010) for a method of facing these problems, we have made no attempt to face them and we have accepted all the links as correct.

In step 1 the top category of domain *d* is looked for in WP^{en}. Let $topCaDom^d$ be this category. Once located $topCaDom^d$ for English, the top categories for the other languages are obtained through the corresponding interlingual links. This is the only step requiring a small amount of human intervention.

In step 2, the initial set of categories is obtained for each language *l* by navigating top down, from the top category, through category/category links, the category graph of WP^l. Although ideally the *WP* category graph is a *DAG*, it is not really the case because two problems: i) the existence of cycles and ii) the presence of backward links.

Both problems have the same origin: the way of building the resource by lots of volunteers working independently. Many cycles

[5] http://dumps.wikimedia.org/

[6] http://www.ukp.tu-darmstadt.de/software/jwpl/

[7] In fact the real top category is "Contents", we have used "Articles" instead as topCat for avoiding that the shortest paths to the top traverse meta-categories.

occur in *WK*, an example, from the Spanish *WP,* is *Drogas → Drogas y Derecho → Narcotráfico → Drogas*. Detecting cycles and removing them is quite straightforward.

The second problem is more serious and difficult to face. When working with English *WP* we discovered that for the domain *Medicine* 90% of the whole *WP* category graph was collected as descendants of the domain top category. Consider the following example, from English *WP: Volcanology → Volcanoes → Volcanic islands → Iceland*. In this case going Top Down from the category *Volcanology* a lot of categories related to *Iceland*, but with no relation with *Volcanology* will be collected. For facing the second problem (backward links) we can take profit of the following information:

- The relative depth of each category *c* regarding $topCaDom^d$, i.e, the length of the shortest path from *c* to $topCaDom^d$.
- The absolute depth of c, computed in the preparatory step, i.e, the length of the shortest path from *c* to *topCat*.
- The absolute depths of *topCat* and $topCaDom^d$.
- The absolute depth of the parent of *c* in the dop down navigation.

We have experimented with several filtering mechanisms, from the very simplest one, pruning the current branch when the depth of *c* is lower than the depth of the parent of *c*, to others more sophisticated. Finally we decided to apply the following filtering: *c* is pruned, and not further expanded, if the relative depth of *c* is greater than the difference between the absolute depths of *topCat* and $topCaDom^d$ plus 1.

Applying this filtering mechanism resulted in reducing the set of involved categories (more than 900,000 without filtering) to a manageable number of 5,874 categories for English.

In step 3 we build the initial set of pages, collecting for each category in the set of initial categories the corresponding pages through the category/page links. The process is, so, quite straightforward. A simple filtering mechanism is performed for removing Named Entities and not content pages.

In step 4, from the two sets built in step 2 and 3 a graph representing the whole set of *TC* for the domain *d* and for all the languages is built. Figure 3 presents an excerpt of this graph. The nodes of the graph correspond to all the pages and categories selected in steps 2 and 3 for all the involved languages. The edges, which are directional, correspond to all the links considered (category → category, category → page, page → category, page → page and interlingual links).

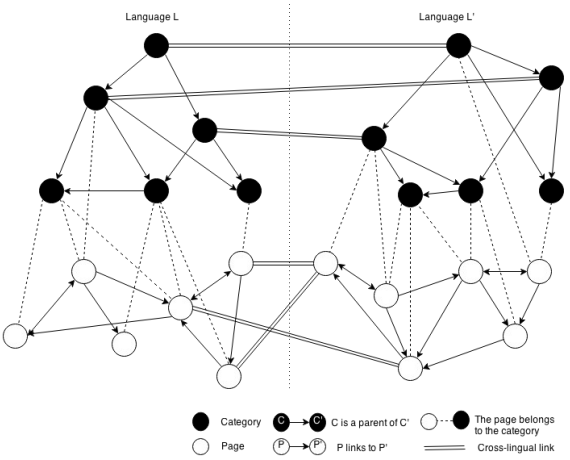

Figure 3: Graph representation of *TC*

In step 5 the nodes of the graph of *TC* are scored. For doing so we use an algorithm inspired in Topic-Sensitive PageRank, (Haveliwala. 2002), in turn based on the original PageRank algorithm, (Page, Brin, 1998).

The original PageRank algorithm is based on a scoring mechanism that allows for a given node upgrading its score accordingly with the scores of its incident nodes. So in this setting all the incident edges are equally weighted and the new score is only affected by the old one and the scores of the incident nodes. As is discussed en section 4, this setting does not work very well and we looked for some form of weighting of the edges, and not only of the nodes for computing the final score of a node.

In the case of nodes corresponding to pages there are three types of incident edges (for nodes corresponding to categories the formulas are similar):

- *il*: inlinks, links from other pages.
- *cp*: links from the categories the page belongs to.
- *ll*: langlinks, links from pages in other languages

The score of a page is computed by adding three weighted addends, one for each type of edge. The formula applied is the following:

$$PR(p) = F_{il} \sum_{il \in inlinks_p} \frac{PR(il)}{L(il)} + F_{cp} \sum_{c \in categories_p} \frac{PR(c)}{L(c)} + F_{ll} \sum_{ll \in langlinks_p} \frac{PR(ll)}{L(ll)}$$

where *PR(i)* is the PageRank score of node *i*, F_t are weights of edges of type *t* (*il, cp*, or *ll*), and L(n) are normalizing factors for pages or categories, computed as:

$$L(p) = F_{ol} \times |outlinks| + F_{pc} \times |cats| + F_{ll} \times |langlinks|$$

for pages and similarly for categories.

Finally, in step 6 the set of nodes corresponding to each language are sorted by descendent score giving the final result of the system. No distinction is made in this sorted sequence between *TC* corresponding to categories and these corresponding to pages.

4 *Experiments and evaluation*

4.1 Initial Settings

We performed some initial experiments for setting the parameters F_t defined in step 5. Finally we set F_{ll} and F_{cp} to 100 and the other parameters to 1. For evaluating these settings we limited ourselves to English and Spanish in the medical domain for which a golden repository of terms, *SNOMED*[8], is available. We consider four scenarios: i) *all_zeroes*, where no scoring procedure is used, ii) *all_ones*, where the standard *PageRank* algorithm is applied, iii) *no_langlinks*, where interlingual links weights are set to zero, i.e. the *TC* for each language are extracted independently, and, iv) *best*, where the setting described above was applied. The results are presented in Figures 4 , for English, and 5, for Spanish. All *PageRank* based scenarios clearly outperform the *all_zeroes* baseline. The differences between these scenarios are small for English but significant for Spanish where *best* outperforms clearly the others.

4.2 Experiments

We applied the procedure described in section 3 to the two domains and 4 languages using the setting of section 4.1. The results are presented in Table 1.

[8] http://www.ihtsdo.org/

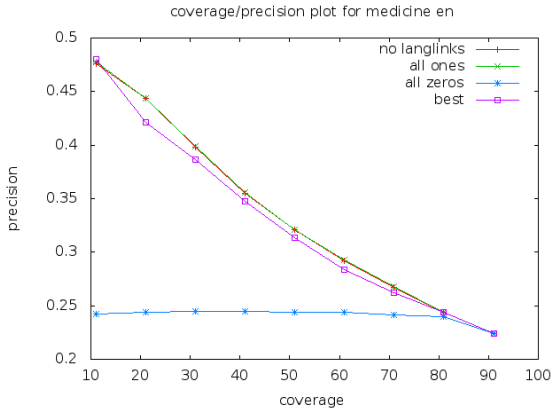

Figure 4: Initial experiments for English

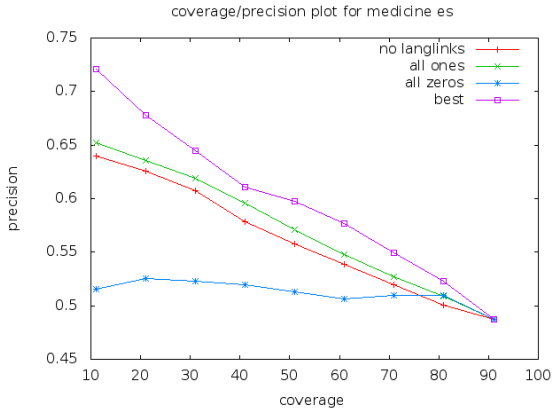

Figure 5: Initial experiments for Spanish

Language	Medicine	Finance
English	67,448	8,711
Spanish	8,872	1,310
Catalan	2,827	674
Arabic	7,318	1,557

Table 1: Overall results of our experiments

The figures in Table 1 are not very informative. Being our system a ranker what is important is accepting as true terms the best ranked until some threshold. We depict, so, in Figures 6 (for medicine) and 7 (for finance) the distribution of *TC* in a coverage/score plots[9]. Content of these Figures and Table 1 are somewhat complementary.

[9] Note that, contrary to Figures 4 and 5 where ordinates display precision, in this case ordinate display scores, i.e. PR values.

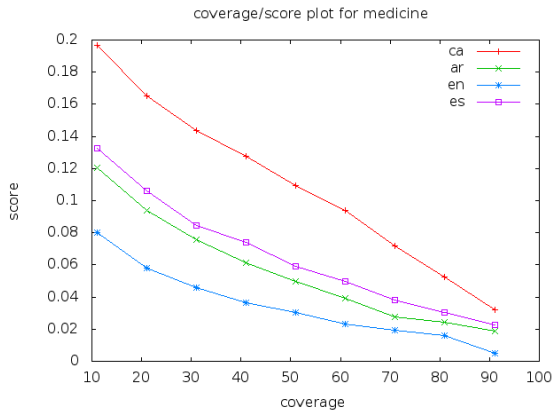

Figure 6: Results for medicine

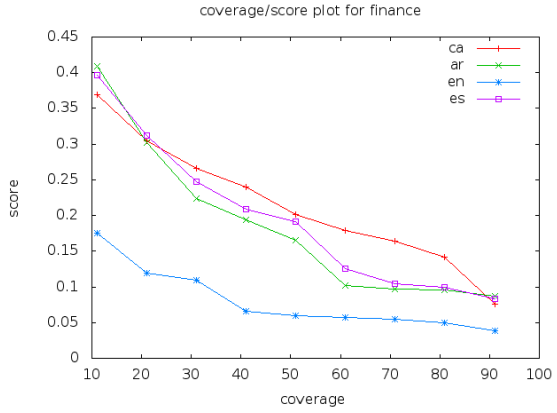

Figure 7: Results for finance

4.3 Evaluation

Evaluation of our results is not easy. For the pairs medicine/English and medicine/Spanish we can use as golden repository *SNOMED* and use as evaluation the results of the *best* curve in Figures 4 and 5. We have measured the correlation between precision in Figures 4 and 5 and score in Figure 6. Pearson's coefficient is 0.93 for Spanish and 0.98 for English, so we are pretty confident on our results for these two pairs. However, as pointed out in (Vivaldi, Rodríguez, 2012), *SNOMED* is far to be a reliable reference, for English only 62% of the correct *TC* were found in *SNOMED*. So the figures in Figures 4 and 5 can be considered a lower bound of the precision. For measuring a more accurate value we performed an additional manual validation[10] over the *TC* not found in *SNOMED* corresponding to the best 20% ranked ones. Figure 8 compares for this rank interval the precisions computed against *SNOMED* golden and those that combines it with the manual evaluation. At can be seen, results improvement is between 20 and 30 points.

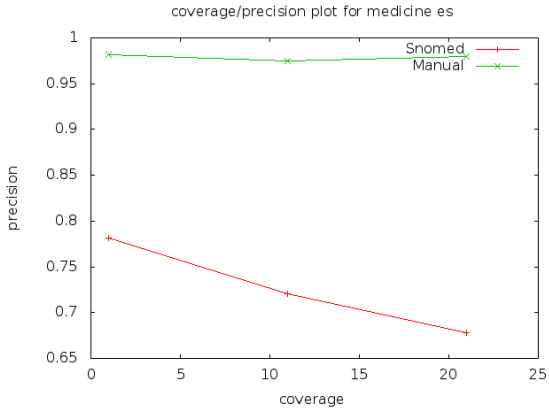

Figure 8: Comparison of SNOMED based and manual evaluation for Spanish

Obviously all these evaluations are in some cases partial and in other cases indirect. A point to be assessed is whether the evaluation results could be extrapolated to other domains and/or languages. For having some insights we computed the Pearson's correlation coefficient between the non-cumulated ranked scores for the different languages. Table 2 shows the results for *medicine*. A very similar result has been obtained for *finance*. The high values of these coefficients seem to support out hypothesis. The score distribution correlates well between all the languages for all the domains. At the beginning of this section we saw that for *medicine* and for the languages English and Spanish scores and precision correlated well too. So our guess is that the evaluation based on *SNOMED* for English and Spanish and the manual one for a segment of Spanish can be likely been extended to the other cases.

A comparison with other systems is not possible globally but we can perform some partial and indirect comparisons with the system closest to ours', (Vivaldi, Rodríguez, 2012). In this work, applied to Spanish and English, one of the domains included is *medicine* and *SNOMED* is used for evaluation. The main differences with ours' are that i) it is a term extractor, not a ranker and, ii) the evaluation is performed over terms belonging to *WordNet*. So the comparison has to be indirect. For the level of precision reported there, 0.2 for English, 0.4 for Spanish, the corresponding coverage in Figures 4 and 5 are 0.8 and 0.9. So, the number of terms we extract are 53,950 and

[10] Performed by the two authors independently, followed by a discussion on the cases with no agreement.

7,985 that clearly outperform largely the 21,073 and 4,083 reported there.

	en	es	ca	ar
en	1.0	0.996	0.990	0.992
es	0.996	1.0	0.995	0.994
ca	0.990	0.995	1.0	0.982
ar	0.992	0.994	0.982	1.0

Table 2: Correlations between non-cumulated ranked scores for the different languages

5 Conclusions and Future work

We have presented a terminology ranker, i.e. a system that provides a ranked list of terms for a given domain and language. The system is domain and language independent and uses as unique Knowledge Source the *Wikipedia* versions of the involved languages. The system proceeds in a cross-lingual way using for scoring a variant of the well-known *PageRank* algorithm.

We have applied the system to four languages and two domains. The evaluation, though not complete, and somehow indirect, and the comparison with a recent system closely related to ours', at least at the level of the source, shows excellent results clearly outperforming the subjects of ours comparisons.

Future work includes i) the application of the system to other domains and, possibly, to other languages and, ii) the improvement of the evaluation setting applying the system to domains for which terminology exists.

No attempt has been made to face the reciprocity and inconsistency of interlingual links. We plan in the near future to analyze these issues and to try to obtain aligned collections of multilingual terminologies.

The software and datasets described in this paper will be made publicly available in the near future through github.

Acknowledgements

The research described in this article has been partially funded by Spanish MINECO in the framework of project SKATER: Scenario Knowledge Acquisition by Textual Reading (TIN2012-38584-C06-01).

We are in debt with three anonymous reviewers whose advices and comments have contributed to a clear improvement of the paper.

References

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

Cabré, M.T., Estopà, R., Vivaldi, J., 2001. Automatic term detection. A review of current systems. *Recent Advances in Computational Terminology* 2:53-87.

Drouin, P., 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1):99-115.

Enguehard, C., Pantera, L., 1994. Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics* 2(1):27-32.

Erdmann, M., Nakayama, K., Hara, T., Nishio, S., 2009. Improving the extraction of bilingual terminology from Wikipedia. *TOMCCAP* 5(4).

Erdmann, M., Nakayama, K., Hara, T., Nishio, S., 2008. An Approach for Extracting Bilingual Terminology from Wikipedia. *DASFAA 2008* 380-392.

Frantzi, K.T., Ananiadou, S., Tsujii, J., 2009. The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. *Lecture Notes in Computer Science* 1513:585-604.

Gabrilovich, E., Markovitch, S., 2009. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research* 34:443-498.

Hartmann, S., Szarvas, G., Gurevych, I., 2012. Mining Multiword Terms from Wikipedia. M.T. Pazienza and A. Stellato: *Semi-Automatic Ontology Development: Processes and Resources* 226-258.

Haveliwala, T.H., 2002. Topic-sensitive PageRank. *Proceedings of the 11th international conference on World Wide Web (WWW '02)* 517-526.

Heid, U., Jauß, S., Krüger, K., Hohmann, A., 1996. Term extraction with standard tools for corpus exploration. Experience from German. *Proceedings of Terminology and Knowledge Engineering (TKE'96)*.

Alkhalifa, M., Rodríguez, H., 2010. Automatically Extending Named Entities coverage of Arabic WordNet using Wikipedia. *International Journal on Information and Communication Technologies* 3(3).

Krauthammer, M.I., Nenadic, G., 2004. Term identification in the biomedical literature. *Journal of Biomed Inform* 37(6):512-26.

Medelyan, O., Milne, D.N., Legg, C., Witten, I.H., 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716-754.

De Melo, G., Weikum, G., 2009. Untangling the Cross-Lingual Link Structure of Wikipedia, *48th Annual Meeting of the Association for Computational Linguistics*.

Niemann, E., Gurevych, I., 2011. The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet}. *Proceedings of the 9th International Conference on Computational Semantics* 205-214.

Page, L., Brin, S., 1998. The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh International Web Conference (WWW-98)*.

Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M., 2005. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *Studies in Fuzziness and Soft Computing* 185:255-279.

Sabbah, Y.W., Abuzir, Y., 2005. Automatic Term Extraction Using Statistical Techniques- A Comparative In-Depth Study and Application. *Proceedings of ACIT'200.*

Sadat, F., 2011. Extracting the multilingual terminology from a web-based encyclopedia. *RCIS 2011* 1-5.

Vivaldi, J., 2001. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD Thesis, Universitat Politècnica de Catalunya.

Vivaldi, J., Rodríguez, H., 2008. Evaluation of terms and term extraction systems. A practical approach. *Terminology* 13(2):225-248. John Benjamins.

Vivaldi, J., Rodríguez, H., 2011. Using Wikipedia for term extraction in the biomedical domain: first experience. *Procesamiento del Lenguaje Natural* 45:251-254.

Vivaldi, J., Rodríguez, H., 2012. Using Wikipedia for Domain Terms Extraction. In Gornostay, T. (ed.) *Proceedings of CHAT 2012: The 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources: co-located with TKE 2012*.

Wolf, E., Gurevych, I., 2010. Aligning Sense Inventories in Wikipedia and WordNet. *Proceedings of the First Workshop on Automated Knowledge Base Construction* 24-28.

Zesch, T., Müller, C., Gurevych, I., 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *LREC 2008: Proceedings of the Conference on Language Resources and Evaluation* 1646-1652.