Using Wikipedia for Domain Terms Extraction

Jorge Vivaldi[1] and Horacio Rodríguez[2]

[1] Universitat Pompeu Fabra, Barcelona, Spain
jorge.vivaldi@upf.edu

[2]Technical University of Catalonia, Barcelona, Spain
horacio@lsi.upc.edu

Abstract. Domain terms are a useful resource for tuning both resources and NLP processors to domain specific tasks. This paper proposes a method for obtaining terms from potentially any domain using Wikipedia.

Keywords: term extraction, domain terminology, Wikipedia

1 Introduction

Even though many NLP resources and tools claim to be domain independent, its application to specific NLP tasks uses to be restricted to specific domains. As the accuracy of NLP resources degrades heavily when applied in environments different from which they were built; a tuning to the new environment is needed.

The basic knowledge sources, KS, needed for performing this tuning are domain restricted corpora and terminological lexicons. The latter is specially challenging and this is the goal of the work described here. Manual acquisition is costly and time consuming due to an extremely low level of agreement among experts [14]. Terminology extraction is more serious in domains in which the distinction between real terms and general words is difficult to establish preventing us of using un-restricted out of domain documents.

In this paper we present an approach for extracting terminological information for a given domain using the Wikipedia (WP) as main KS. It is domain/ language independent, we have applied it to two languages (Spanish and English) and to some randomly chosen domains. In section 2 we introduce both term extractions and WP. Then, in section 3 and 4 we present both our approach for obtaining the terminologies and its evaluation. Finally, in section 5 we present some conclusions and future work.

2 State of the art

Terms are usually defined as lexical units that designate concepts of a thematically restricted domain. As shown in [2] and [10], many methods have been proposed to extract terms from a corpus. Some of them are based on linguistic knowledge, like in [6]. Others use statistical measures, such as ANA [4]. Some approaches combine both linguistic knowledge and Statistics, such as [3] or [5]. A common limitation of most extractors is that they do not use semantic knowledge, therefore their accuracy is limited. Notable exceptions are Metamap [1] and YATE [11].

WP is the largest on-line encyclopaedia; its information unit is the *Page* that basically describes a concept. The set of pages and their links in WP form a directed graph. A page is assigned to one or more WP categories in a way that categories can be seen as classes linked to pages. At the same time, a category is linked to one or more categories structuring themselves too as a graph. WP has been largely used as KS for extracting valuable information ([8]).

3 Our approach

In previous works we developed two alternative methods for extracting terminology for a domain using WP categories and pages as KS. The aim is to collect these units from WP such that their titles could be considered terms of the domain.

The first approach ([13]) follows a top down strategy starting in a manually defined top category for the domain. The problem of this approach was its limited recall due to the absolute dependence of the extracted term candidates on such category.

The second ([14]) follows a bottom up strategy. It starts with a list of TC, obtained from some domain specific text. In this approach both precision and recall are affected: i) the TC set is reduced to the list and ii) requires a top category that conditions the process as in the first approach.

In this paper we propose to combine both approaches to overcome these limitations. For accessing WP we have used Gurevych's JWPL [15]. Scaling up our methodology implies four additional not independent tasks over the work done previously, namely: i) choosing an appropriate domain taxonomy; ii) selection of category tops corresponding to the domains considered; iii) obtaining an initial set of TCs and iv) allowing a neutral automatic evaluation.

As domains taxonomy we use Magnini's Domain Codes, *MDC* [7]. Such codes enrich WordNet[1]. We can use WN for a cheap, though partial, evaluation of our method.

Our claim is that our method could be applied to any language owning a relatively rich WP. However, the results presented in this paper are reduced to English and Spanish and a randomly[2] selected subset of MDC consisting of 6 domains is presented and discussed. Figure 2 presents the overall process, it is organized into 8 steps (step 6 is iterated until convergence). The overall process is repeated for the two languages and domains involved (Agriculture, Architecture, Anthropology, Medicine, Music and Tourism). From now on let *lang* be the language considered and *dc* the Magnini's domain code, in *MDC*. The first step of our method consists of extracting from the WN corresponding to the language *lang* all the variants contained in all the synsets tagged with domain code *dc*. This results on our first set of TC, $terms_0$.

The second step consists of mapping *dc* to a set of WP categories. First we look whether dc occurs in the WP category graph (CG). If it is the case (it is true for 90% of *dc* for English), the set {dc} is selected. Otherwise we look if *dc* occurs in the WP

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

[2] Medicine has been included for allowing an objective evaluation, as reported in section 4.

page graph (PG). If this case we obtain the categories attached to the page. Otherwise a manual assignment, based on an inspection of WP is performed. The step results on an initial set of categories *categories₀*.

categories₀ contains mostly a unique category but when it has been built from a page it can contain noisy categories. In the third step *categories₀* is cleaned by removing neutral categories and categories attached to domain codes placed above *dc* in MDC taxonomy.

The basis of our approach consists of locating two subgraphs, *CatSet* in CG, and *PageSet* in PPG having a high probability of referring to concepts in the domain, our guess is that the titles of both sets are terms of the domain.

Step 4 builds the initial set of categories, CatSet_0, expanding the tops. Starting in the top categories of dc, CG is traversed top down, avoiding cycles, performing cleaning as in step 3[3]. The categories in this initial set are scored, using only the links to parent categories, as shown in formula (1), then all categories with scores less than 0.5 are removed from the set resulting in our initial set, CatSet_0, as shown in Figure 2.

$$score_{cat} = \frac{\left|parents_{cat}^{ok}\right|}{\left|parents_{cat}^{ok}\right| + \left|parents_{cat}^{ko}\right|} \qquad (1)$$

$parents_{cat}^{ok}$, $parents_{cat}^{ko}$: set of parents categories under/outside domain tops

In step 5 the initial set of pages, $\mathrm{PageSet}_0$, is built. From each category in CatSet_0 the set of pages, following category-page links, is collected in $\mathrm{PageSet}_0$. Each category is scored according to the scores of the pages it contains and each page is scored according both to the set of categories it belongs to and to the sets of pages pointing to/from it. Three thresholding mechanisms are used: Microstrict (accept a category if the number of member pages with positive score is greater than the number of pages with negative score), Microloose (similarly with greater or equal test), and Macro (using the components of such scores, i.e. the scores of the categories of the pages). Formula (2) formalizes the scoring function.

$$score_{page} = comb(score_{pag}^{ocats}, score_{pag}^{input}, score_{pag}^{output}) \qquad (2)$$

where

$$score_{cpagt}^{ocats} = \frac{\sum_{\forall cat \in cats(page)} score_{cat}}{\left|cats(page)\right|}$$ with cats(page)= set of categories of page

$$score_{pag}^{input} = \frac{\sum_{\forall p \in input(page)} score_{p}}{\left|input(page)\right|}$$ with input(page)= set of pages of pointing to page

$$score_{pag}^{output} = \frac{\sum_{\forall p \in output(page)} score_{p}}{\left|output(page)\right|}$$ with output(page)= set of pages pointed from page

and *comb* is a combination function of their arguments

Then, in step 6, we iteratively explore each category. This way the set of well scored pages and categories reinforce each other. Less scored categories and pages are

[3] WCG was preprocessed for attaching to every category the depth in the categories taxonomy.

removed from the corresponding sets. As seen in (2) and (3), a combination function is used to compute a global score of each page and category from their constituent scores. Several voting schemata have been tested. We choose a decision tree classifier using the constituent scores as features. A pair of classifiers, *isTermcat* and *isTermpage*, independent of language and domain, were learned. The process is iterated, leading in iteration i to $CatSet_i$, $PageSet_i$, until convergence[4]. All the sets $CatSet_i$ and $PageSet_i$, are collected for all the iterations for performing the following step.

$$score_{cat} = comb(score^{strict}_{cat}, score^{loose}_{cat}, score^{micro}_{cat}) \qquad (3)$$

where

$$score^{strict}_{cat} = \frac{count_{\forall page\equiv pages(cat)}\left(score_{page} > 0.5\right)}{\left|pages(cat)\right|}$$

with *pages(cat)* = set of pages of cat

$$score^{loose}_{cat} = \frac{count_{\forall page\equiv pages(cat)}\left(score_{page} \geq 0.5\right)}{\left|pages(cat)\right|}$$

$$score^{micro}_{cat} = \frac{\forall page \in pages(cat)}{\left|pages(cat)\right|}$$

and $comb$ is a combination function of their arguments

In step 7 a final filtering is performed for selecting from all the $CatSet_i$ and $PageSet_i$ corresponding to all the iteration the one with best F1. According to the way of building these sets (in step 6) it is clear that precision increases from one iteration to the following at a cost of a fall in recall, as some TC are removed in each iteration Before computing F1 both category and pages sets are merged into a unique term candidate set for each iteration (there are more elements in $PageSet_i$ than in $CatSet_i$ and the intersection of both sets is usually not null. Finally, we evaluate the results as shown in section 4.

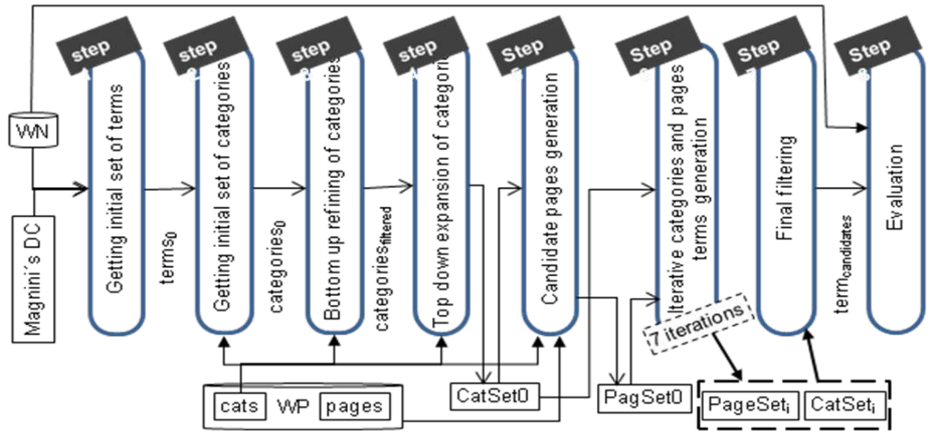

WN
Magnini´s DC
step 1
Getting initial set of terms
terms0
step 2
Getting initial set of categories
categories0
step 3
Bottom up refining of categories
categoriesfiltered
step 4
Top down expansion of categories
Step 5
Candidate pages generation
Step 6
Iterative categories and pages terms generation
7 iterations
Step 7
Final filtering
termcandidates
Step 8
Evaluation
cats
WP
pages
CatSet0
PagSet0
PageSeti
CatSeti

Fig. 1. Methodology

[4] In all the cases, convergence was reached in less than 7 iterations.

4 Evaluation

Evaluation of a terminology is a difficult task ([14]) due to a) the difficulty in doing it through human specialists, b) the lack/incompleteness of electronic reference resources and c) disagreement among them (specialists and/or reference resources).

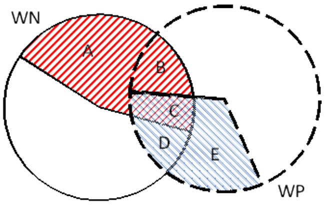
For this reason, we set two scenarios for evaluation. In the first one we analyze the results of Medicine for which we use SNOMED[5] as gold standard. In the second one, as we lack references our evaluation is only partial. Our thought is that the results in the Medicine domain related can be extrapolated to the others domains.

We use for comparison two baseline systems, one based on WN (Magnini) and the other based on the alignment of WN senses to WP pages in NG, [9].

Magnini baseline consists simply on, giving a domain code, *dc*, of Magnini's taxonomy, collecting all the synsets of WN assigned to *dc*, and considering as TCs all the variants related to these synsets. This approach has the obvious limitation of reducing coverage to the variants contained in WN; also it is rather crude because no score is attached to TCs, despite their degree of polisemy or domainhood.

NG map WP pages with WN synsets reaching a 0.78 F1 score. Our baseline is built collecting all the synsets corresponding to *dc* and from them all the WP pages aligned with the synset.

In the first scenario, the set of obtained TCs is compared with the two baselines for English and with the first one for Spanish and with the SNOMED repository. In the second scenario (covering the other domains) the comparisons are reduced to baselines. For both evaluations we need to consider the information shown in Figure 2[6].



$$\text{Precision} = \frac{C}{C+D} \qquad \text{Recall} = \frac{C}{B+C}$$

- A: WN domain variants not found in WP;
- B: WN domain variants found in WP but not considered in the domain by Magnini;
- C: WN domain variants found in WP;
- D: WN variants belonging to the domain according the WP but not according WN;
- E: WP pages/categories belonging to the domain but not found in WN;
- A+B+C: WN variants for a given domain;
- C+D+E: WP pages/categories discovered.

Fig. 2. Terms indirect evaluation

As shown in Figure 2, our system starts from the set of WN variants defined by [8], as belonging to the domain. Then it finds a number of WP pages and categories. Some of them are included in the set of variants already defined by Magnini but it also discovers new TC in WP. The evaluation can only be done using the terms already defined by Magnini and assuming their correctness. It is expected that terms discovered in WP will have similar precision values.[7]

[5] A comprehensive repository of Spanish/English terminology. See http://www.ihtsdo.org/

[6] The figure reflects Magnini's baseline, reflecting Niemann_Gurevych's is similar.

[7] Magnini assignment has been done in a semiautomatic way; therefore, they are not error free.

Using the sets of terms defined in Figure 2 we calculate the corresponding precision/recall values shown in Table 1. For each language and domain the initial number of WN variants and the precision/recall values are presented. As mentioned above such values are calculated against information obtained from the Magnini's domains. The table include also the results obtained using SNOMED.

Table 1. Results of the experiments (* at the best F1 values, ** evaluated using SNOMED-CT)

Domain		Tourism		Architecture		Music	
Language		EN	ES	EN	ES	EN	ES
Terms in WN	Total	744	441	303	143	1264	747
	In WP	554	286	244	112	1035	567
Precision [%]*	Cat.	33.33	100.00	0.00	85.71	50.57	50.00
	Page	15.65	85.71	36.59	59.52	11.11	27.42
Recall [%]*	Cat.	0.36	0.70	0.00	5.36	4.25	1.94
	Page	4.15	2.10	6.15	22.32	6.37	3.00
New Terms		1061	42	122	189	7046	614

Domain		Agriculture		Anthropology		Medicine			
Language		EN	ES	EN	ES	EN	EN**	ES	ES**
Terms in WN	Total	396	209	1106	651	2451		1595	
	In WP	238	137	909	443	1783		954	
Precision [%]*	Cat.	7.14	20.00	24.49	60.00	47.64	100.00	72.48	100.00
	Page	6.10	10.94	5.16	25.93	19.86	100.00	40.53	100.00
Recall [%]*	Cat.	0.42	0.73	1.32	0.68	10.21	6.56	11.32	16.25
	Page	10.50	5.11	5.06	1.58	16.32	9.76	15.93	54.51
New Terms		1491	193	6100	973	7855	3541	2225	2413

Table 2. Comparison of the results for Medicine/English among different approaches

Approaches	EWN	SNOMED	Precision	Recall
Ours	450	279	62.00	42.02
Magnini	1257	664	52.82	100.00
NG	190	150	78.95	22.59

A first consideration to be taken into account in analyzing the results shown in Table 1 is the own characteristics of WP as a source of domain terms. In particular:

- CG may change across languages. See for example Medicine and Veterinary. Although definitions are similar in both Spanish and English WPs, the former considers both as siblings whilst the latter considers it as a subcategory of former. This difference causes a large difference in the TC direct/indirect linked to them;
- English WP is a densely-linked resource; this causes unexpected relations among TC. Consider for example the domain "Agriculture" and the terms "abdomen" or "aorta". Both TCs are considered to be related to the domain due to a link among "Agriculture" → "veterinary medicine" which may be considered wrong;
- WP is an encyclopaedic resource; therefore, the termhood of some TC may be controversial. See for example: "list of architecture topics" in Architecture.

Low recall shown in Table 1 is due to the way of computing it, relating to terms in both WN/ WP. So, most of the extracted terms do not account for recall, eg, for tourism in English 1061 terms are extracted but only 25 of them occurs both in WN/WP. Due to the difficulties in the evaluation of the term lists, the characteristics of MDC and WP we perform additional evaluation for some domains. The results for Tourism were evaluated manually by the authors and the results for Medicine has been evaluated using SNOMED. Below we describe and analyse such additional evaluations.

1. Tourism (Spanish). We performed a manual evaluation of the TCs proposed. Partial evaluation takes as reference the list of EWN variants found in WP although, such variants not always are considered by WP to belong to the domain. Therefore it is possible to perform such evaluation taking into account this fact. It has been performed in two different ways for DC thresholds values ranging from 0 to 0.2:
   i) Precision/recall calculation: recall rises from 1.7 to 50%.
   ii) Error ration calculation: error rate decreases 70.96% to 0%.
2. Medicine. The use of SNOMED allows a better evaluation. The results show a considerably improvement in the precision/recall values (see Table 1, columns tagged with ** and Table 2). Magnini's offers the highest score in recall because the terms considered are all under its *dc* (ie. B in Fig. 2 is null). NG obtains the best score in precision with a low recall. Our results are in the middle.
3. Nevertheless there are some problems in using this repository such as:

   - Complex term: Some terms in this database are coordinated terms. See for example the Spanish TC: *enfermedades hereditarias y degenerativas del sistema nervioso central* (genetic and degenerative disorders of the central nervous system). It causes that none of the coordinated term are detected.
   - Some entries exist only as specialized. See for example the Spanish TC *glándula* (gland), it only exists as a more specialized terms like *glándula esofágica* (esophageal gland) or *glándula lagrimal* (lacrimal gland).
   - Number discrepancies among a WP category and the related SNOMED entry.
   - Missing terms like: *andrología* (andrology) or *arteria cerebelosa media* (medial cerebellar artery), present only in WP snapshot used for this experiment.
   - The results for Medicine and English are low. It is due to the number of entries, in our version, is much lower than those for Spanish (852K *vs* 138K).

5 Conclusions and future work

In this paper we present a new approach for obtaining the terminology of a domain using the category and page structures of WP in a language/domain independent way. This approach has been successfully applied to some domains and languages. As foreseen the results evaluation is a difficult task, mainly due to issues in the reference list. Also the encyclopaedic character of WP conditioned the list of new terms obtained. The performance may also change according the domain/language considered.

The current definition of domain (a set of WP categories) could be problematic when considering subdomains or interdisciplinary domains (like law, environment or information science). This will be a topic for future research/improvement.

In the future we plan to improve the final list of terms by: i) improve the exploration of the WP in order to reduce the false domain terms, ii) using the WP article text as a factor of pertinence of a page, iii) a better integration of both exploration procedures and iv) enlarge the number of proposed TC by using interwiki information.

Acknowledgement

This research has received support from the projects KNOW2 (TIN2009-14715-C04-04) and "RicoTerm 4" (FFI2010-21365-C03-01). We also thank the anonymous reviewers for their comments and suggestions.

References

1. Aronson A., Lang F.: An overview of MetaMap: historical perspective and recent advances. *JAMIA 2010* 17, p. 229-236 (2010).
2. Cabré M.T., Estopà R., Vivaldi J.: Automatic term detection. A review of current systems. *Recent Advances in Computational Terminology* 2, p. 53-87 (2001).
3. Drouin P.: Term extraction using non-technical corpora as a point of leverage. Terminology 9(1), p. 99-115 (2003).
4. Enguehard C., Pantera L.: Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics* 2(1), p. 27-32 (1994).
5. Frantzi K. T., Ananiadou, S., Tsujii, J.: The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. LNCS, Volume 1513, p. 585-604 (2009).
6. Heid, U., Jauß, S., Krüger K., Hofmann, A.: Term extraction with standard tools for corpus exploration. Experiece from German. In Proceedings of TKE'96. Berlin (1996).
7. Magnini B., Cavaglià G.: Integrating Subject Field Codes In WordNet. In 2nd LREC (2000).
8. Medelyan, O., Milne, D., Legg C., Witten, I. H.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies 67 (9), p. 716-754 (2009).
9. E. Niemann, Gurevych I.: The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In: Proceedings of the 9th International Conference on Computational Semantics, p. 205-214 (2011).
10. Pazienza M.T., Pennacchiotti M., Zanzotto F.M.: Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *StudFuzz* 185, Springer-Verlag, p. 225-279 (2005).
11. Vivaldi J.: Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD Thesis, Universitat Politècnica de Catalunya (2001).
12. Vivaldi J., Rodríguez H.: Evaluation of terms and term extraction systems: A practical approach. *Terminology* 13(2), p. 225-248 (2007).
13. Vivaldi J., Rodríguez H.: Finding Domain Terms using Wikipedia. In 7th LREC (2010).
14. Vivaldi J., Rodríguez H.: Using Wikipedia for term extraction in the biomedical domain: first experience. In *Procesamiento del Lenguaje Natural* 45, p. 251-254 (2010).
15. Zesch T., Müller C., Gurevych I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In 6th LREC p. 1646-1652 (2008).