Finding Domain Terms using Wikipedia

Jorge Vivaldi[1], Horacio Rodríguez[2]
[1]Institute for Applied Linguistics, Roc Boronat 138, UPF, Barcelona, Spain
[2]Software Department, Jordi Girona Salgado 1-3, UPC, Barcelona, Spain
E-mail: jorge.vivaldi@upf.edu, horacio@lsi.upc.edu

Abstract

In this paper we present a new approach for obtaining the terminology of a given domain using the category and page structures of the Wikipedia in a language independent way. The idea is to take profit of category graph of Wikipedia starting with a top category that we identify with the name of the domain. After obtaining the full set of categories belonging to the selected domain, the collection of corresponding pages is extracted, using some constraints. For reducing noise a bootstrapping approach implying several iterations is used. At each iteration less reliable pages, according to the balance between on-domain and off-domain categories of the page, are removed as well as less reliable categories. The set of recovered pages and categories is selected as initial domain term vocabulary. This approach has been applied to three broad coverage domains: astronomy, chemistry and medicine, and two languages: English and Spanish, showing a promising performance. The resulting set of terms has been evaluated using as reference those terms occurring in WordNet (using Magnini's domain codes) and those appearing in SNOMED-CT (a reference resource for the Medical domain available for Spanish).

1. Introduction and motivation

Since the 80s there was an acute need, from different disciplines and goals, to automatically extract terminological units from specialized texts. Computational linguists, applied linguists, translators, interpreters, scientific journalists and computer engineers have been interested in automatically isolating terminology from texts for a number of purposes: building of glossaries, vocabularies and terminological dictionaries; text indexing; automatic translation; building of knowledge databases; improving automatic summarization systems, construction of expert systems and corpus analysis. Typical approaches involve linguistic and/or statistical systems with results not fully satisfactory (see Cabré et al., 2001 for a revision). One of the reasons of this behaviour is that none of first approaches use semantic knowledge.

In Vivaldi et al. (2002) we faced the problem of automatically extracting domain terminology using Domain Markers (DM), with WordNet (WN), Fellbaum, 1999, as Knowledge Sources. We defined a DM as a WN or EuroWordNet[1] (EWN, Vossen, 2004) entry (a synset) whose attached strings belong to the domain, as well as the variants of all (or at least most of) its descendents through the hyponymy relation. In that initial research, DMs were selected manually starting with a set of seed words for the domain, looking for the corresponding synsets in WN and exploring their environment. As this procedure is costly and difficult to scale up, in Vivaldi et al. (2004) we faced the problem of automatically selecting the DM. The basic Knowledge Source was in this case a glossary of initial terms for the domain. So, a fatal loop was found: For extracting terminology (in our approach) it was necessary a set of DM that in turn needed an initial terminology for being extracted. In this paper a new approach is presented that tries to break the loop using an external Knowledge Source, Wikipedia, for providing the initial set of terms for the domain, which triggers the whole procedure.

Wikipedia[2] (WP), is by far the largest encyclopaedia in existence with more than 3 million articles in its English version (EWP) contributed by thousands of volunteers. WP experiments an exponential growing. There are versions of WP in more than 200 languages although their coverage (number of articles and average size of each article) is very irregular.

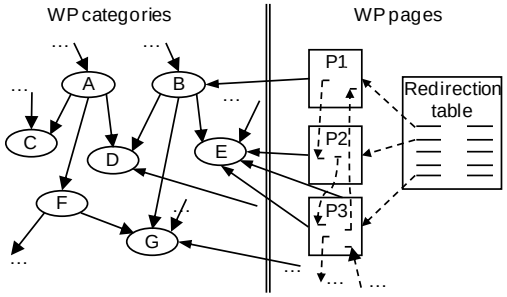

Figure 1: The graph structure of WP

WP information unit is the "Article" (or "Page"). Internally, an article may contain links to other articles in the same language by means of "Article links". There are about 15 output article links (links are not bidirectional) in average in each WP article. The set of articles and their links in WP form a directed graph.

The whole article is assigned to one or more WP categories (through "Category links") in such a way that categories can be seen as classes that are linked to pages (belonging to the category). At the same time, a category is linked to one or more categories (super and sub categories) structuring themselves as classes that are also organized as a graph (see Zesch, Gurevych, 2007, for an interesting analysis of both graphs). In Figure 1 we can

[1] We applied our method to English and Spanish.

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

see an overall image of both connected graphs. This bi-graph structure of WP 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 (see, for instance, Suchanek (2008)). 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.

Besides article and category links, WP pages can contain "External links", which point to external URLs, and "Interwiki links", from an article to a presumably equivalent, article in another language. There are in WP several types of special pages: "Redirect pages", i.e. short pages which often provide equivalent names for an entity, and "Disambiguation pages", i.e. pages with little content that links to multiple similarly named articles.

While edges between categories usually (but not always) have a clear semantics (hypernymy, hyponymy), edges between pages lack tags or explicit semantics. Also, some categories are added to WP by convenience for structuring the database or due to its encyclopaedic character (e.g. "scientists by country", "Chemistry timelines" or "Astronomical objects by year of discovery" among many others). Other categories are used temporally for monitoring the state of the page (e.g. "All articles lacking sources", "Articles to be split" ...), we name these categories "Neutral Categories". Due to such facts it becomes difficult, just navigating through its structure, to discover which entry belongs to which domain.

WP has been extensively used for extracting lexical and conceptual information: Ponzetto, Strube, 2008 and Suchanek, 2008, build or enrich ontologies from WP, Milne *et al.*, 2006 derive domain specific thesauri, Atserias et al, 2008 produce a semantically annotated snapshot of EWP, Medelyan *et al.*, 2008, Mihalcea, Csomai, 2007, and Wu *et al.*, 2007 perform semantic tagging or topic indexing with WP articles. Closer to our task are the works of Toral *et al.*, 2006 and Kazama, Torisawa, 2007 which use WP, particularly the first sentence of each article, to create lists of named entities. Relatively low effort has been devoted to exploit the multilingual information of WP. Ferrández *et al.*, 2007, Richman, Schone, 2008 and, more recently, Erdmann *et al.*, 2008 are notable exceptions. See Medelyan *et al.* (2009) for an excellent survey.

Extracting information from WP can be done easily using a Web crawler and a simple HTML parser. The regular and highly structured format of WP pages allows this simple procedure. There are, however, a lot of APIs providing easy access to WP online or to the database organized data obtained from WP dumps[3]. Some interesting systems are Waikato's WikipediaMiner toolkit[4], U. Alicante's wiki db access[5], Strube and Ponzetto's set of tools[6], Iryna Gurevych' JWPL[7], etc. We have used this later resource for our research.

The proposed system should be language and domain independent. Therefore, for any language to be considered the only limitation, regarding both quality and quantity, depends only of the WP for such language.

The key idea of our approach is using the category graph of WP starting with a top category that we identify with the name of the domain. From this top, we extract the set of (presumably) relevant categories, traversing the graph following sub-category links. For avoiding noise we apply rigid constraining on the categories to visit. From the set of categories selected, the collection of corresponding pages is extracted, also under some constraints. For reducing noise a bootstrapping approach implying several iterations is used. The set of recovered pages and categories is selected as initial domain term vocabulary.

After this introduction the organization of the paper is as follows: Section 2 presents an overview about the research already done in this area. Section 3 shows with some details our approach to this issue while section 4 and 5 presents our experiments and evaluation results respectively. Finally in section 6 we will derive some conclusions and proposals for future work.

2. State of the art

Bernardini *et al.* (2006) propose the Wacki system, a method for extracting both corpora and terminology for a domain. The approach is recall-oriented and so not useful for our purposes.

(Magnini *et al.*, 2000) have enriched WN with domain information on the basis of a general classification that includes 164 domains/subdomains (structured in a rather flat taxonomy). Following a semiautomatic procedure, one or more domain tags have been assigned to each synset.

(Montoyo *et al.*, 2001) propose a way of enriching WN with about 30 IPTC subject codes. Their approach follows the Specification Marks Method, previously used for Word Sense Disambiguation tasks. Also (Buitelaar, *et al.*, 2001) propose a method for domain specific sense assignment using GermaNet together with a set of relevance measures. A closely related task is the automatic extraction of domain ontologies from general ones using domain corpora. (Missikoff *et al.*, 2002) present an interesting approach.

In an automatic term extraction system, applied to the medical domain, (Vivaldi, Rodríguez, 2002) use DM, as defined before. About 50 borders were manually identified and used as a basis for term extraction. In (Vivaldi, Rodríguez, 2004) it is showed how public

[3] http://en.wikipedia.org/wiki/Wikipedia_database
[4] http://wikipedia-miner.sourceforge.net/
[5] http://www.dlsi.ua.es/~atoral/
[6] http://www.eml-research.de/english/research/nlp/download/
[7] http://www.ukp.tu-darmstadt.de/software/jwpl/

available vocabularies may be used to enrich EWN with domain information in a fully automatic way.

3. Methodology

As said above, the basic idea of our method consists of given a domain name (e.g. "Computing") to find it in WP as a category, to obtain the full set of domain term candidates (*DTC*) belonging to such domain. Such set of *DTC*'s will be a list of all the categories and page titles that our system considers that belong to the domain of interest.

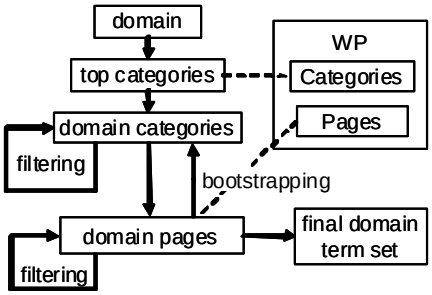

Figure 2: Overview of the method

Choosing the right top for the domain is a crucial issue in our approach. Usually, the name of the domain corresponds to a category in WP (it was the case of the three domains faced in our experiments) otherwise looking for the right top[8] is performed manually. For reaching our objectives we start looking for all subcategories and pages related to such domain. From such lists we remove all proper names and service classes[9]. Then, we recursively explore each category and repeat the same process again. See Figure 2 for an overview of the approach. Proceeding without any additional check has the inconvenience that it is relatively easy to add names belonging to a different domain. Consider the following example: assuming we wish to obtain all terms for "Computing" domain, we may arrive to the WP entry for "semantics". There are several paths to reach this category from the top, as "computing → theoretical computer science → semantics" or "computing → software → software engineering → formal methods → semantics" among others. The problem is that, proceeding forward in this way, we may reach entries like "lexical semantics" or even "weak pronoun" which clearly do not belong to the target domain. The problem is really serious. In the case of the topic "Chemistry", for English, recovering the categories related by the (supposed transitive) subcategory relationship with the category "Chemistry" resulted on 188,374 categories. Obviously most of these categories do not belong to the domain and have to be removed for avoiding such amount of noise. Another problem is the presence of cycles. In the case of "Chemistry" 247 cycles were detected (for instance, 'Oil pipelines' and 'Natural gas pipelines'). The problem of cycles is that two categories involved in a cycle can prevent each other for being removed by the procedures described below. In order to avoid such situations, we act at both category and page levels.

At category level the procedure is rather straightforward. We proceed in two steps:

We extract all the descendent, avoiding loops, of the top categories with no constraint. Let *CatSet1* this set. All the cycles are detected and collected in this step.
For each category $c \in CatSet1$ we count the number of direct super-categories $\in CatSet1$ and the corresponding $\notin CatSet1$. Neutral Categories are not taken into account for these counts. If the first count is lower than the second we remove the category. If the category c is involved in a loop, detected in step 1, with one of its super-categories this later super-category is not taken into account. We also remove those categories that we consider proper names. At this point we consider a category to be a proper noun when it is a polylexical unit and all its components have an initial upper case. We iterate step 2 until convergence.
This lead to *CatSet2* that is included in the final set of DTC's

At page level the procedure is not so simple. We can filter out both pages and categories, using in this case the scores of the pages assigned to the category. We define the WP domain coefficient (WPDC) as follows:

$$WPDC(dtc) = \frac{\sum_{\forall c \in termCats} PathToDomainCat(c)}{\sum_{\forall c \notin termCats} PathToDomainCat(c)} \quad (1)$$

where:

dtc:	domain term candidate
termCat:	set of WP categories associated to *dtc*
PathToDomainCat(c):	= 1 if $c \in CatSet2$; = 0 otherwise

For filtering out pages the idea is, for a given DTC, to obtain from WP all its assigned categories, *termCat*. Then, for each of such categories we check if they belong to *CatSet2*. Additionally, we define a threshold for WPDC (1 in our experiments, as a rather conservative approach) and only the DTC's whose WPDC is higher that such threshold is allowed to survive and therefore are added to the target list and used for additional iterative explorations. It should be taken into consideration that a *DTC* may correspond to a WP category as well as a WP page.

Proceeding in this way in the above example, the domain term candidate "semantics" has been assigned 5 categories ("Linguistics", "Philosophy of language", "Semiotics", "Theoretical computer science" and "Philosophical logic"). From such category just one reach the target domain; therefore *WPDC(semantics)*=0.25 and consequently the search is pruned at this point.

[8] In some domains more than one top are considered.

[9] Classes defined by WP management people for internal organization (eg. "Wikipedia stubs", "Wikipedia cleanup", "Wikipedia CD Selection", etc.).

For filtering out categories we can use the scores of the pages belonging to each category. For a category *cat* let *catTerm* the set of pages associated to it. In our approach we chose to build three different ways to evaluate *cat* and combine all the evaluation results through a voting mechanism that perform the final decision. Such evaluation methods are the following:

- MicroStrict. Accept *cat* if the number of elements of *catTerm* with positive scoring is greater than the number of elements with negative scoring.
- MicroLoose. Similarly with greater or equal test
- Macro. Instead of counting the pages with positive or negative scoring we use the components of such scores, i.e. the number of categories associated with the elements of *catTerm* belonging or not to *CatSet2*

The above method may be repeated iteratively several times in order to improve the results. The results reached are exemplified in Table 1.

#	DTC	Micro Strict		Micro Loose		Macro		Vote	Result
		ok	ko	ok	ko	ok	ko		
1	electroquímica (electrochemistry)	13	5	16	2	36	12	+3	Accept
2	quesos (cheeses)	0	8	6	2	8	12	-1	Reject
3	óxidos de carbono (carbon monoxide)	1	1	2	0	4	3	+2	Accept

Table 1: Examples of filtering for the domain 'Chemistry' in Spanish

4. Experiments and Evaluation

In order to evaluate the above mentioned methodology several tests has been performed. First at all it should be considered the difficulty to evaluate the results produced by the proposed tool. As pointed out in Vivaldi *et al.* (2008), the evaluation of a terminology is a difficult task mainly due to the lack/incompleteness of reference resources to be used as test bed and, often, the disagreement among them. In this research, we are looking for a domain/language independent tool which makes the task even more difficult. For such reason we chose two different evaluation procedures:

1. partial evaluation for two domains: Chemistry and Astronomy. It was done using the set of DM produced by our methods with the proposal of (Magnini et al., 2000). It does not include all DTC but just 1960 entries whose unique domain is "chemistry" (1672 nouns) and 416 (319 nouns) whose domain list includes "chemistry". It must be considered that WN and EWN are general purpose resources; therefore, its coverage for the chosen domain/language is low (25% for Chemistry and 15% for Astronomy). Thus, some common terms (as "mol" and "sunspot" among many others) are not included.
2. full evaluation for Medicine. In this case we use SNOMED-CT, a reference resource for this domain. It is a structured collection of medical terms that has a wide coverage of the clinical domain. The terms are organized in a number of hierarchies where nodes are linked using both hierarchical and non hierarchical ("causative agent", "finding site" and "due to") relations. Since 2002 a Spanish edition is published regularly. The release used in this experiment is dated on October 2009 and contains more than 800k entries[10].

Anyway, at first we applied a manual preliminary evaluation on a first set of terms we obtained (for the category terms corresponding to the Spanish-Chemistry pair). Two independent evaluators examined the set. This set contained 114 categories. The agreement between the two evaluators was relatively high (kappa over 80%). After discussing the conflictive cases (all of them belonging to borders between Chemistry and very close domains (Medicine, Physics) the result was of considering 84 cases as valid terms of the domain (74%) and 30 cases as erroneous ones. With these positive initial tests we went into a more serious evaluation.
The next two sections will show the results obtained using the above mentioned evaluation procedures.

4.1 Partial evaluation

Table 2 shows the results obtained for Chemistry and Astronomy domains for English and Spanish using different selection methods defined in the proposed methodology. The huge number of initial categories in some cases, obviously containing a lot of noise, is due to the occurrence of links pointing to upper categories. The results shown in this table are limited to the initial categories and the results after the first iteration. Figures 3 and 4 show the evolution of the precision over several iterations.

Domain			Chemistry		Astronomy	
Language			EN	ES	EN	ES
Initial Categories			188374	2070	188816	44631
#Categories after pruning			1334	557	790	143
Iteration #1	Categories		49	43	5	6
	Precision		93,9	62,8	0	16,7
	Pages found	Loose	833	1038	284	119
		Strict	580	700	284	81
	Prec. [%]	Loose	61,3	52,6	34,8	31,9
		Strict	62,7	56,6	37.2	27,2

Table2. Results of the experiments

As foreseen, the number of pages obtained with "loose" method is larger that those found using the "strict" method. Correspondingly, the precision score reached is higher for the latter method. See for example Chemistry (English), the strict method found 580 pages (from which 364 belongs to the correct domain, therefore the precision is 62.7%) against 833 (precision: 61.3%) for the loose method. Regarding the difference among both procedures they seem to perform according our previsions. The latter proposes some erroneous terms like "characteristic",

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

"congo red", "elixir of life", "interior", "ordinal number" or "neon lamp" among others. At the same time, some valid terms are wrongly discarded ("oxytocin", "sulphonamide", "chemical structure"). This is because in such cases the number of support categories is the same of non support categories.

A result that requires some clarification is those obtained for Astronomy (English): only five categories were found but none of them, according the reference chosen, belong to this domain. This not true, because at least three of them are correct: "astrodynamics", "astrometry" and "celestial mechanics". The reason is that Magnini classifies such entries as "factotum" (a subject field code to collect synsets very ambiguous or difficult to classify). The result is almost identical for Spanish.

The results for pages in English WP for Chemistry also show some discrepancies. For example the term "resonance" is discarded from this domain because it is considered as belonging to "physics" or "music". This happens because there a sense missing in WN: "resonance", in chemistry: it designates a key component of valence bond theory. Other times it happens that there are terms belonging to two domains like "pyrite" (it is term used in geology but it also is a chemical compound —FeS_2—) or "magnesium hydroxide" and "menthol" that are chemical compound but are also used in medicine.

Also there are some mistakes due to the encyclopaedic character of the WP. This is the case of "photocathode" that appears related to "electrochemistry" while Magnini classifies it as belonging just to electricity.

It should be noted also that among the terms not found in WN (3,258 after the first iteration) there are many that clearly are used in Chemistry ("carbon trioxide", "chemical transformation", "chlorosulfuric acid" or "heavy metals", among others) while others are dubious ("experimental value", "global field" or "group representation") or at least their specialized character is uncertain.

The results obtained for both domains and languages are quite similar. This fact confirms our claim about the difficulty of evaluating the results. There are some decisions taken by Magnini's proposal that are questionable and cause a number of reported errors: some terms belong to two domains (e.g. medicaments or body substances are medical terms but also chemical substances) but they are systematically classified in just one of them. It should also be taken into account that Magnini's proposal is relatively old as it was made on WN 1.6[11] and outdated version of this resource. Other errors are caused by the encyclopaedic character of WP; therefore, the pruning/filtering procedures need to be improved.

[11] WN had foreseen a domain classification of the entries. Unfortunately this information has not been included. The work of Magnini is the only domain information currently available for WN/EWN.

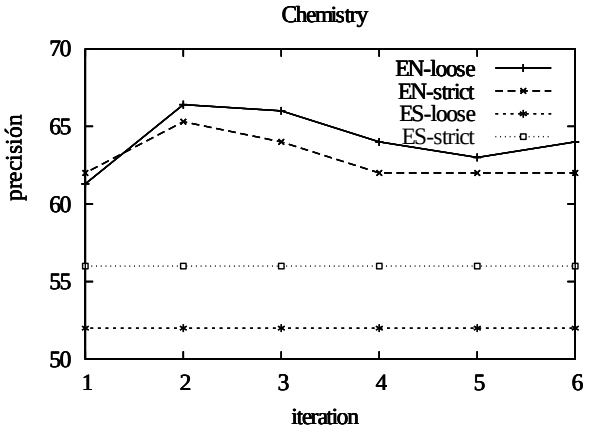

Figure 3. Results for several iterations (Chemistry)

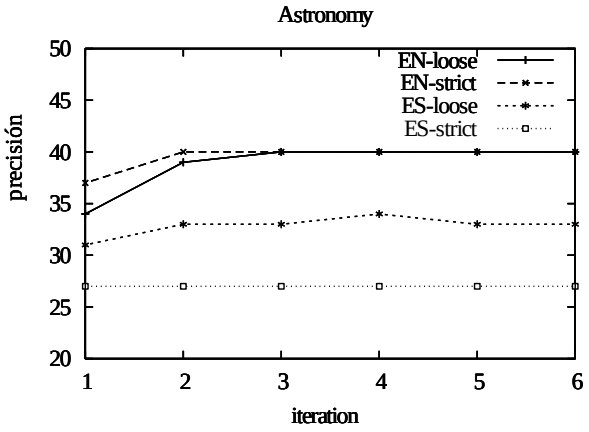

Figure 4. Results for several iterations (Astronomy)

4.2 Full evaluation

Table 3 shows the results obtained for Medicine using the Spanish WP. The precision figure reached using SNOMED-CT, compared with those obtained using WN as a reference, show a strong enhancement. It is clearly due to the improvement in the list of reference terms used in the evaluation. The difference among SNOMED-CT and WN is greater than two orders of magnitude.

Evaluation using			WN	SNOMED-CT
Initial Categories			2431	
Categories after pruning			839	
Iteration #1	Categories		174	394
	Precision		27,6	54
	Page	Loose	2091	4182
		Strict	1724	3492
	Prec. [%]	Loose	21,0	58
		Strict	23,2	62

Table3. Results of the experiments for Medicine (Spanish)

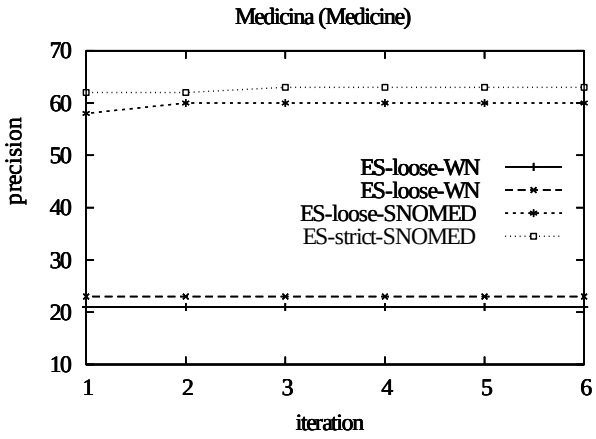

Figure 5. Results for several iterations (Medicine)

In spite of the specialized character of the evaluation resource, it accepts as valid term sequences like "whisky", *puro* ("cigar"), *tortura* ("torture"), *ubre* ("udder") and *fuego* ("fire") but no accept others like *oral cancer* that clearly seems to be a medical term. The point here is that whisky and other beverages are in the "substances" subtree of SNOMED-CT probably because they are the cause of a number of diseases.

Others terms are rejected due to minor differences: *enfermedades del sistema digestivo* ("gastrointestinal tract diseases") versus *enfermedades del sistema digestivo* ("gastrointestinal tract disease") or simply by missing like *espina ilíaca antero-superior* ("anterior superior iliac spine"), *medicina intensiva* ("intensive-care medicine") or *fisiopatología* ("pathophysiology").

We observe that there are some systematic errors as for example the inclusion of Named Entities. Figure 6 shows a typical example: the entity "Campari" is considered as belonging to the domain due to the existence of a path ("alcoholic beverages by country" → "alcoholic beverages" → "alcohol" → "psychotropics") to the domain name node. The appropriate way to cutting this path is detecting that "alcoholic beverages by country" is not a taxonomic link. Cutting these links can be done by simple pattern matching techniques.

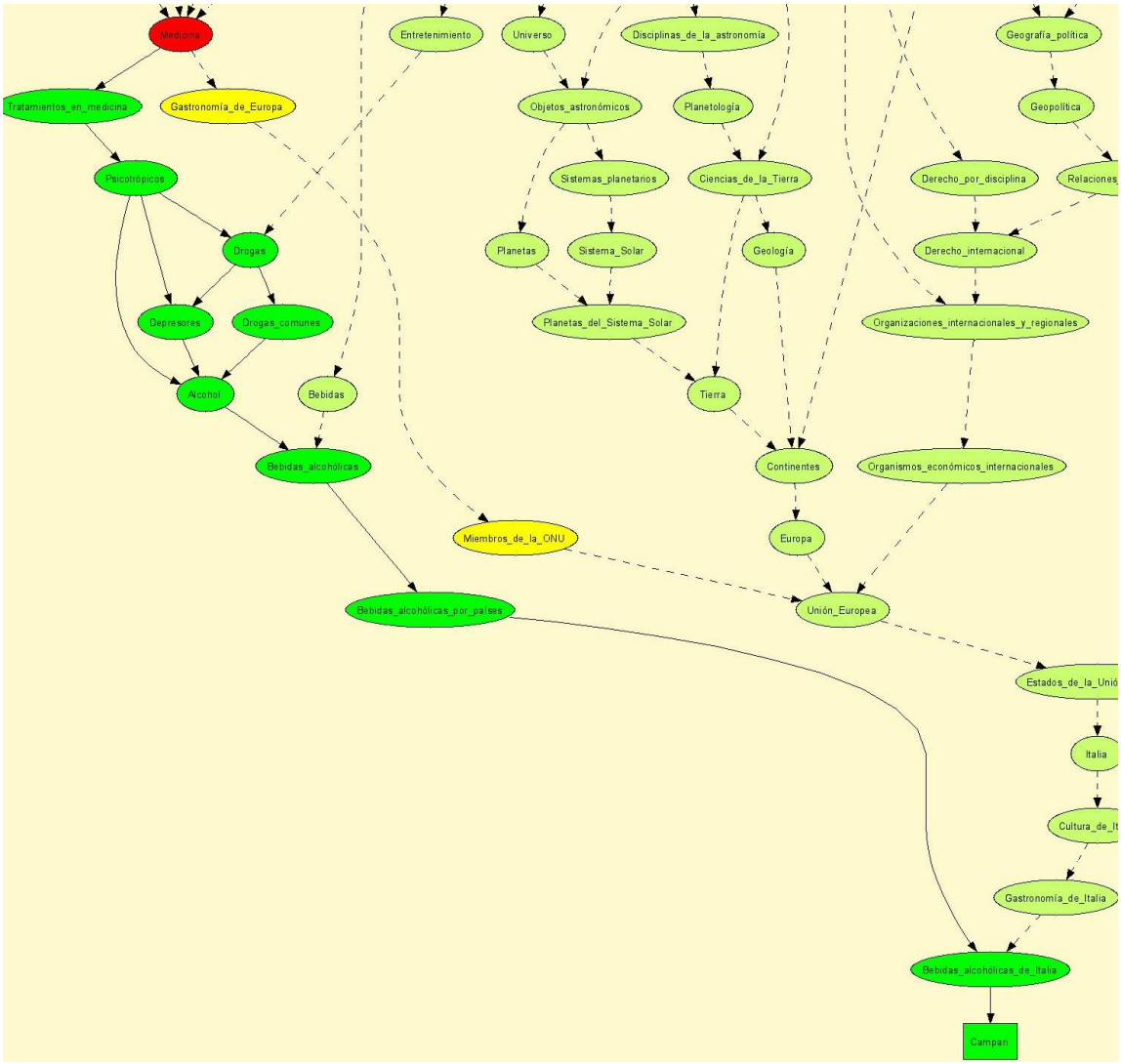

Figure 6. Example of Named Entity inclusion in the *Medicina* (Medicine) domain.

5. Conclusions

In this paper we present a new approach of obtain the terminology of a domain using the category and page structures of WP in a language independent way. This approach has been applied to some domains and languages showing a promising performance.

Using the same list of words, the evaluation results show a clear improvement when using a specialized and complete resource in the domain (SNOMED-CT instead of WN). In our opinion this fact shows that quality of the results for Chemistry and Astronomy are better than those showed by the evaluation. The problem is to obtain a good enough resource to perform the evaluation. Another problem is that an outdated WN was used to evaluate the results. Some of the problems commented in the evaluation section (like terms belonging to two domains, ex. "medicines" designates to both chemical and pharmaceutical compounds) are solved with newer versions on WN.
As mentioned above, the filtering procedure shows some drawbacks (as the inclusion of Named Entities in the term list) that could be improved by cutting some category links.

In the next future we plan to apply our method to other languages, other domains and other reference resources (using domain glossaries and/or any kind of reference term list). Using not only the categories and pages titles as proposed terms for the domain but also the vocabulary contained in the best scored pages is another line of research we are currently exploring.

6. Acknowledgements

This research has received support from the project KNOW2 (TIN2009-14715-C04-04), from the Spanish Ministry of Science and Innovation. We also thank three anonymous reviewers for their comments and suggestions.

7. Bibliography

Atserias, J. Zaragoza, H. Ciaramita M. and Attardi. G. (2008). "Semantically Annotated Snapshot of the English Wikipedia". Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).

Bernardini, S.; Baroni M. and Evert S. (2006). "A WaCky Introduction". Wacky! Working papers on the Web as Corpus, Bologna: Gedit. 9-40.

Cabré, M. T.; Estopà, R. y J. Vivaldi (2001) "Automatic Term Detection: A Review Of Current Systems". In Bourigault, D.; Jacquemin C. y M.C. L'Homme (eds.) *Recent Advances in Computational Terminology*. John Benjamins Publishing Company. Amsterdam.

M. Erdmann, K. Nakayama, T. Hara, S. Nishio (2008): Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia, in IPSJ Journal of Information Processing (Jul. 2008).

Fellbaum, C. (ed.) (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Ferrández, S. Toral, A. Ferrández, O. Ferrández, A. Muñoz R. (2007) Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering. NLDB 2007: 352-363

Kazama, J. and Torisawa, K. (2007) Exploiting Wikipedia as External Knowledge for Named Entity Recognition. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning

Magnini B. and G. Cavaglià, (2000). "Integrating Subject Field Codes In WordNet". In Proceedings of the 2nd LREC International Conference. Athens, Greece

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

Medelyan, O. Witten, I. H. Milne D. (2008) Topic indexing with Wikipedia. In Proc of Wikipedia and AI workshop at the AAAI-08 Conference. Chicago, US

Mihalcea, R. Csomai A. (2007) Wikify!: linking documents to encyclopedic knowledge CIKM 2007: 233-242

Milne, D., Medelyan, O. and Witten, I.H. (2006) Mining Domain-Specific Thesauri from Wikipedia: A case study. In Proc IEEE/WIC/ACM International Conference on Web Intelligence, WI'06, pp. 442-448, Hong Kong, China,.

Ponzetto, P; Strube, M. (2008). "WikiTaxonomy: A large scale knowledge re source". In: Proceedings of the 18th European Conference on Artificial Intelligence, Patras, Greece, 21-25 July, 2008 pp. 751-752.

Richman, A.. and Schone, P. (2008) Mining Wiki Resources for Multilingual Named Entity Recognition. In Proceedings of ACL-08

Suchanek F. (2008) Automated Construction and Growth of a Large Ontology PhD-Thesis. Max-Planck-Institute for Informatics. U. Saarbrücken, Germany

Toral, A. Muñoz. R. (2006) A proposal to automatically build and maintain gazetteers for Named Entity Recognition using Wikipedia. In Proceedings of the Workshop on New Text, 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento (Italy). April 2006.

Vivaldi, J. and Rodríguez, H, (2002). "Medical Term Extraction using the EWN ontology". In Proceedings of Terminology and Knowledge Engineering (pp 137-142). Nancy.

Vivaldi J. y H. Rodríguez (2004) "Automatically selecting domain markers for terminology extraction". In: Proceedings of the 4th LREC International Conference. Pp. 1729-1732.

Vossen P. (2004) EuroWordNet: a multilingual database of autonomous and language specific wordnets connected via an Inter-Lingual-Index. International Journal of Lexicography, Vol.17 No. 2, OUP, 161-173.

Wu, F. Hoffmann, R. Weld D. (2007) Autonomously Semantifying Wikipedia, In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM-07), Lisbon, Portugal, November, 2007.

Zesch, T. Gurevych, I. (2007) Analysis of the Wikipedia Category Graph for NLP Applications. In Proceedings of the (Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, 2007)