Using Wikipedia to Validate the Terminology Found in a Corpus of Basic Textbooks

Jorge Vivaldi[1], Luis Adrián Cabrera-Diego[2], Gerardo Sierra[2], María Pozzi[3]

[1]Institut Universitari de Lingüística Aplicada, UPF. [2]Instituto de Ingeniería, Universidad Nacional Autónoma de México. [3]Centro de Estudios Lingüísticos y Literarios, El Colegio de México.

[1]Roc Boronat 138, 08018 Barcelona, Spain. [2]Torre de Ingeniería, Basamento, Av. Universidad 3000, 04510, Mexico City, Mexico. [3]Camino al Ajusco 20, Pedegral de Santa Teresa, 10740, Mexico City, Mexico.

E-mail: jorge.vivaldi@upf.edu, lcabrerad@iingen.unam.mx, gsierram@iingen.unam.mx, pozzi@colmex.mx

Abstract

A scientific vocabulary is a set of terms that designate scientific concepts. This set of lexical units can be used in several applications ranging from the development of terminological dictionaries and machine translation systems to the development of lexical databases and beyond. Even though automatic term recognition systems exist since the 80s, this process is still mainly done by hand, since it generally yields more accurate results, although not in less time and at a higher cost. Some of the reasons for this are the fairly low precision and recall results obtained, the domain dependence of existing tools and the lack of available semantic knowledge needed to validate these results. In this paper we present a method that uses Wikipedia as a semantic knowledge resource, to validate term candidates from a set of scientific text books used in the last three years of high school for mathematics, health education and ecology. The proposed method may be applied to any domain or language (assuming there is a minimal coverage by Wikipedia).

Keywords: automatic term recognition, Wikipedia, corpus, public knowledge repositories usage.

1. Introduction

A scientific vocabulary is a set of terms designating scientific concepts. Such vocabulary can serve as input to many applications that range from the production of dictionaries to improving NLP systems. In our case, we are interested in identifying the basic scientific vocabulary (BSV) in Mexican Spanish as a multi-purpose linguistic resource for research. For this, we set up the Corpus of Basic Scientific Texts in Mexican Spanish (COCIEM) containing selected textbooks used to teach science and mathematics to 6 to 18 year old students.

Following the work presented in Cabrera-Diego et al. (2011), this paper proposes an improved method to obtain and validate the BSV of Mexican Spanish. This method has been successfully applied to several domains (mathematics, health education and ecology, all from high school) in a domain-independent way.

After this introduction, we firstly present some previous work in this field, we then introduce our methodology. Next, we describe the validation process, results and some issues found. Finally, we present our conclusions and propose some future work.

2. Related work

As shown in Cabré, Estopà & Vivaldi (2001) and Panzienza, Pennacchiotti & Zanzotto (2005), there are several methods to extract terms from a corpus. These can be classified according to whether they are based on:

- linguistic knowledge, like Heid (Heid et al., 1996);
- statistical measures, such as ANA (Enguehard & Pantera, 1994); and
- a combination of both linguistic knowledge and statistical measures, i.e. hybrid methods, for example Termext (Barrón-Cedeño et al., 2009) or TermoStat (Drouin, 2003).

Only a few of them use semantic knowledge, e.g. TRUCKS (Maynard & Ananiadou, 2000), YATE (Vivaldi, 2001) and MetaMap (Arson & Lang, 2010)[1], to validate their resultant terms. The use of semantic knowledge allows the quality of the results to be improved. A common characteristic of those systems is that they are domain and resource oriented and consequently, their adaptation to other knowledge fields is costly and time consuming.Therefore, a new approach is necessary to reach our target.

3. Methodology

The first step for identifying the BSV was to obtain a list of term candidates (TC) from the COCIEM using YATE. The second step consisted of two parallel processes: a) manual validation of the set of term candidates by specialists, and b) the set of term candidates was analysed using Wikipedia. The last step consisted in the comparison of both sets of results for evaluation purposes. The full methodology proposed is shown in figure 1. In the following subsections, each component of the system is described in detail.

3.1 Corpus

COCIEM is a compilation of the most widely used textbooks in Mexico for physics, chemistry, biology, mathematics, health education and ecology. This set of books includes theoretical and practical knowledge that correspond to the current scientific curricula for each school year. The aim was to incorporate a truly

[1] Strictly speaking Metamap is not a term extractor but a concept mapper. It means that a given term candidate string is mapped to a set of concepts. In the case of Metamap, such string is mapped to a UMLS concept, while our tool simply shows that such string may be the lexicalization of a concept of the domain, without referring to any specific set. As well, in contrast to other similar tools we use Wikipedia as a semantic resource for validating the term candidates.

representative set of textbooks of all levels of pre-university education.

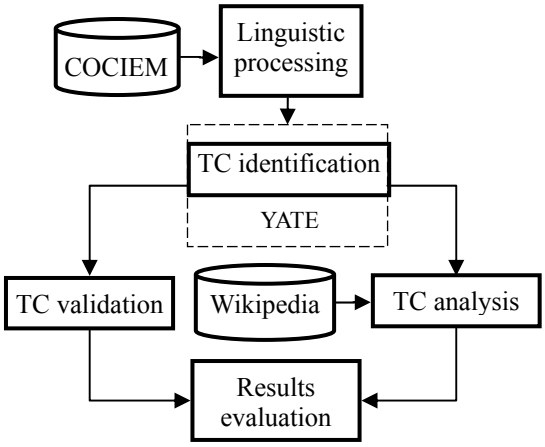


Figure 1: Diagram of the project's methodology

Specifically, COCIEM consists of 92 textbooks (3.6M tokens) classified in three different levels which in turn, are classified into scientific subjects as follows:

- Elementary School (0.3M tokens): natural sciences[2] and mathematics;
- Junior High (2.0M tokens): biology, mathematics, ecology, physics and chemistry; and
- High School (1.3M tokens): biology, mathematics, health education, chemistry, physics and ecology.

3.2 Automatic identification of term candidates

To improve the identification of term candidates we used a term extractor based on linguistic knowledge and not only on statistical measures as in Cabrera-Diego et al. (2011). More specifically, we applied the extraction module defined in YATE. This was decided based on the possibility of having more accurate and complete lists of term candidates.

The input to the extraction module is the corpus text with basic linguistic processing: phrase and token[3] segmentation, morphological analysis and Part-of-Speech tagging. On the basis of such information, this module selects those text sequences that satisfy some specific patterns; for this paper we considered the following: noun, noun-adjective and noun-preposition-noun. The approach we followed is similar to the splitting module presented in Bourigault (1994). Thus, a given unit should both start with a noun and cannot comprise pronouns, adverbs, conjunctions and verbs. Also, YATE can be configured to accept only a given set of prepositions.

3.3 Validation of term candidates

One of the reasons for the lack of good results in the term recognition field is that most systems do not use semantic information to validate their results. A promising alternative is the use of encyclopaedias as knowledge resources. Currently, the obvious choice is Wikipedia, the largest free, multidomain, multilingual encyclopaedia. There are versions in more than 270 languages although the coverage is very irregular. This term validation task has been done by processing a dump as described in Zesch, Müller & Gurevych (2008).

As shown in figure 2, Wikipedia is organised into two connected graphs: the *category graph* and the *page graph*. On the one hand, the category graph is organised as a taxonomy where each category may be connected to an arbitrary number of super/sub categories (often considered as hyperonym/hyponym links). On the other hand, articles are linked between them forming a directed graph. Both graphs are connected together because every article is assigned to one or more Wikipedia categories; see Zesch & Gurevych (2007).

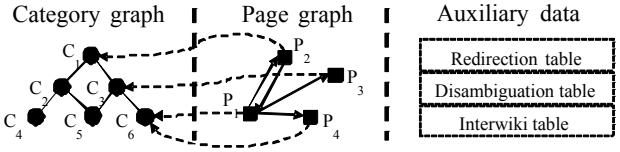


Figure 2: Wikipedia's internal organisation

Wikipedia's bi-graph structure is far from being error-proof:

- categories assigned to a page do not always denote the categories to which the article belongs;
- links between categories do not always indicate hyperonymy/hyponymy relationships;
- several organisation schemes coexist in the category graph; and
- some categories are used for structuring Wikipedia or for monitoring pages.

It therefore becomes rather difficult, just by navigating through the structure, to discover which entry belongs to which domain. In spite of these difficulties, Wikipedia has been extensively used in NLP applications (see Mendelyan et al., 2009 for details and references).

Some applications using Wikipedia's category graph have become especially relevant, for example: large scale taxonomy induction (Ponzetto & Strube, 2011), domain taxonomy building (Kotlerman et al., 2011), ontology building in cooperation with other resources (Suchanek, et al., 2008) and computing semantic relatedness (Ponzetto & Strube, 2007 and Milne & Witten, 2008).

Regarding the accuracy and completeness of Wikipedia articles, it has been favourably compared against Encyclopaedia Britannica, at least in the biochemical domain by Luyt et al. (2007). It has also been successfully compared in Milne et al. (2006) with a specialised domain thesaurus like Agrovoc[4]. The conclusion has been that it covers 50% of the terms, in particular the ones most widely used.

In spite of that, due the collaborative nature of Wikipedia its development is asymmetrical, which means that evolution and completeness of every domain may be different reflecting the interest of resource editors.

[2] Natural sciences include topics related to biology, ecology and the environment, physics, health education and anatomy.

[3] In this case a token can be a string of one or more words or a part of a word.

[4] See: http://aims.fao.org/website/AGROVOC-Thesaurus/sub

The full procedure to evaluate a term candidate is similar to those used by YATE and EuroWordNet (Vivaldi, 2001). It starts by defining the domain of interest as one or more Wikipedia categories; we name these *domain borders*. Usually each domain border coincides with the domain name (e.g.: "Economics") but sometimes it is necessary to use more than one Wikipedia category to define a domain. For example: "Ecology" requires a number of Wikipedia categories ("Environment", "Climatology", etc.).

From this point and for every term candidate, the procedure consists in:

- find a Wikipedia page corresponding to such term candidate:
  - in case of redirection, the redirected term is considered to be the original term[5];
  - in case of ambiguity, such ambiguity is solved using both the domain coefficients (see below) and the distance to the domain border. At the end, it takes the disambiguated term as the original term[6];
- find all Wikipedia categories associated to that page;
- recursively explore the category graph following all super categories links found in the previous step until the domain border or Wikipedia top is reached; and
- sort the list of term candidates according their termhood.

We use the information collected during this exploration to define several *domain coefficients* to calculate the termhood of term candidates (i.e. the association degree of the term candidates with the domain). The calculation of the domain coefficient of a given term candidate *t* is based on the following formulas:

- number of paths

$$DCnc(t) = \frac{NP_{domain}(t)}{NP_{total}(t)}$$

where:

$DCnc(t)$:domain coefficient based on the number of paths

$NP_{domain}(t)$: number of paths to the domain borders

$NP_{total}(t)$: number of paths to the top[7]

- length of paths

$$DClc(t) = \frac{LP_{total}(t) - LP_{domain}(t)}{LP_{total}(t)}$$

where:

$DClc(t)$: domain coefficient based on the length of paths

$LP_{domain}(t)$: length of paths to the domain borders

$LP_{total}(t)$: length of paths to the top[8]

- average length of paths

$$DClmc(t) = \frac{ALP_{total}(t) - ALP_{domain}(t)}{ALP_{total}(t)}$$

where:

$DClmc(t)$: domain coefficient based on the average length of paths

$ALP_{domain}(t)$:average length of paths to the domain borders

$ALP_{total}(t)$: average length of paths to the top[9]

In the case that $LP_{domain}(t)$ and $LP_{total}(t)$ or $ALP_{domain}(t)$ and $ALP_{total}(t)$ were equal, the value of DC*lc*(t) or DC*lmc*(t) is 1, since the entire paths that go to the top pass through the domain borders.

These formulas are based on Vivaldi & Rodríguez (2011) while in Cabrera-Diego et al. (2011) the formulas presented in Vivaldi & Rodríguez (2010) were used.

The value of the domain coefficient ranges from 0 (none of the paths go through the domain border) to 1 (all the paths go through the domain border). Figure 3 shows an example of the calculation of such coefficients for the term *blood*.

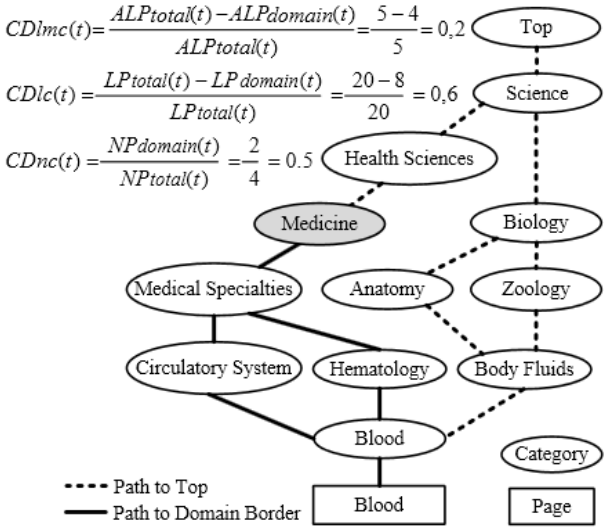


Figure 3: Example of calculation of domain coefficient for the term *blood*

Figure 4 shows a portion of the Spanish Wikipedia category graph corresponding to the ambiguous term (only some paths go through the domain border) "variable dependiente" (*dependant variable*) while figure 5 shows the unambiguous term (all the paths go through the domain border) "derivada" (*derivative*).

[5] A correction coefficient is used in the final rank.

[6] Ibid.

[7] The paths to the top are counted just to the domain borders if they pass through them.

[8] The paths that go to the top but traverse the domain borders are measured until the domain borders.

[9] Ibid.

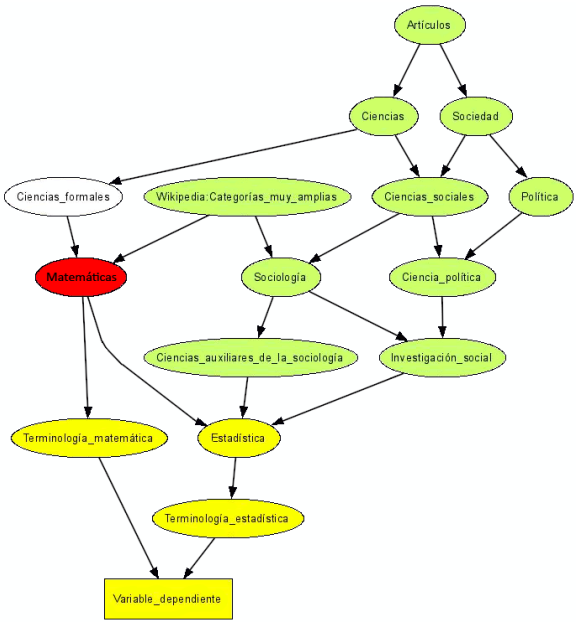


Figure 4: Graph for the term *variable dependiente* (dependant variable)

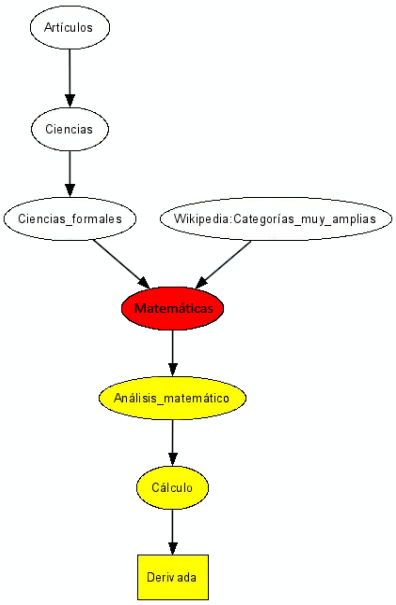


Figure 5: Graph for the term *derivada* (derivative)

4. Results and evaluation

The domain coefficient values resulting from the evaluation reflect the termhood of the candidates according to Wikipedia. They may be classified into four groups:

- $DC(t) = 1$. Term candidate clearly belongs to the domain (see figure 5 for an example).
- $0 < DC(t) < 1$. Term candidate is used in several domains (see figure 4 for an example). Usually, the higher the value the stronger the domain relation is.
- $DC(t) = 0$. Term candidate is not related to the domain.
- $DC(t) = -1$. Term candidate is not found in Wikipedia.

The results have been evaluated using precision and recall measures. For this purpose, the resulting term candidates were evaluated as follows:

- Mathematics: university students with knowledge of terminology;
- Ecology: a biologist trained in terminology with some domain dictionaries[10]; and
- Medicine: using SNOMED-CT [11], a well-known medical terminology resource.

It should be noted that the evaluated term candidates and therefore the evaluation itself have been done over the list of term candidates resulting from different extraction processes instead of those actually occurring in the full corpus. This step was mandatory due to the size of the documents under consideration. See Vivaldi & Rodríguez (2007) for a discussion about evaluation of term extraction systems.

Figure 6 shows the results obtained from the evaluation of our system in the Health Education domain while figure 7 and figure 8 represent the evaluation for the domains of Mathematics and Ecology, respectively.

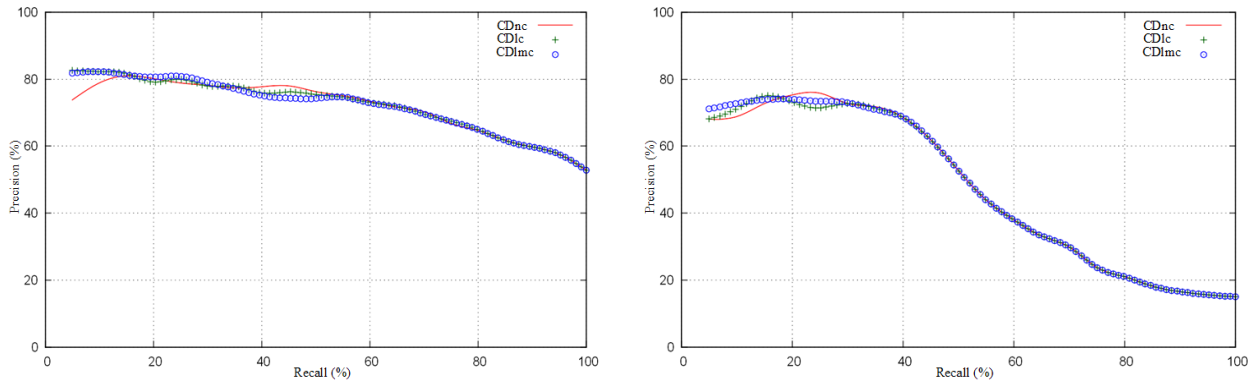
The evaluation of each domain is split into four precision-recall curves, corresponding to the three patterns analysed and the last one corresponds to the combination of all three of them. Compared to other previously published results for similar systems, we consider our results to be reasonably good. In each case, our system reaches high precision with relatively high values of recall, and the vast majority of the term candidates are very well ranked. Also, in the cases where the term candidate is ambiguous in Wikipedia, the disambiguation process works properly. See for example the cases of *suma* (addition), *valor* (value) or *triángulo* (triangle).

In spite of the characteristics of Wikipedia the results are similar to those obtained by similar tools as reported in Vivaldi & Rodríguez. (2007).

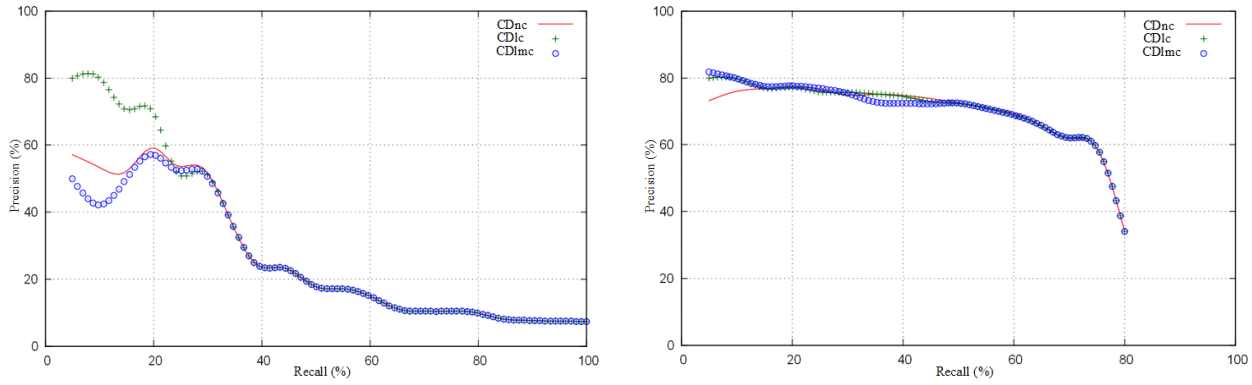
Comparing the results obtained for this paper to those presented in Cabrera-Diego et al. (2011) it must be noted that the term candidate selection is very different. At that time such candidates were obtained using statistical methods with a frequency threshold while now they have been obtained by linguistic analysis. The candidate selection reported in this paper is better than the one presented in Cabrera-Diego et al. (2011) because it has less false candidates, thanks of the use of syntactic patterns, and it includes low frequency candidates; however, in this case, there are some errors caused by the linguistic processing resources, such as: a) POS tagger/lemmatisation module due tagger errors, lack/size of context or dictionary missing entries and b) proper name detection module that may fail in some circumstances. Therefore, the selection of term candidates may fail in some cases for causes that are outside term extraction control.

[10] The first dictionary was *Diccionario de ecología: paisajes, conservación y desarrollo sustentable para Latinoamérica* (Sarmiento, 2001); the second one was *Diccionario Ecológico* (http://www.peruecologico.com.pe/glosario.htm).

[11] Systematized Nomenclature of Medicine - Clinical Terms. See: http://www.ihtsdo.org/ for details.

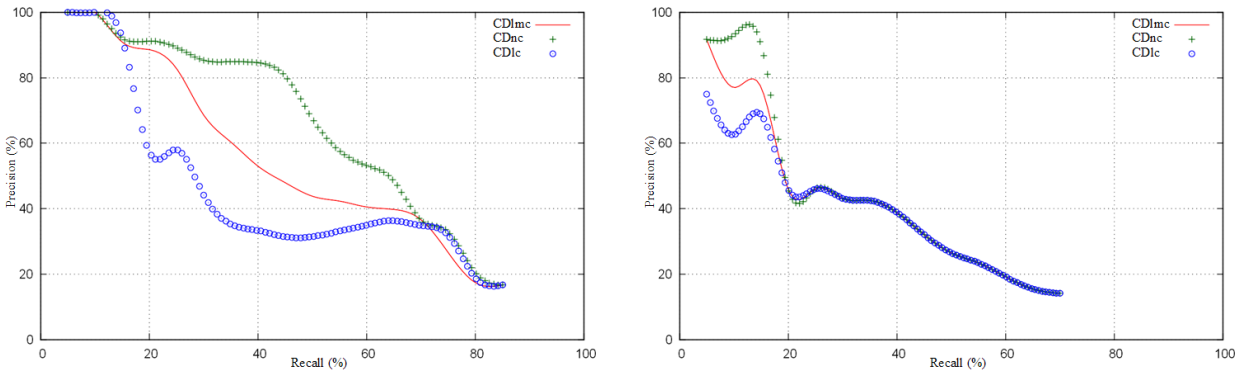


a) b)

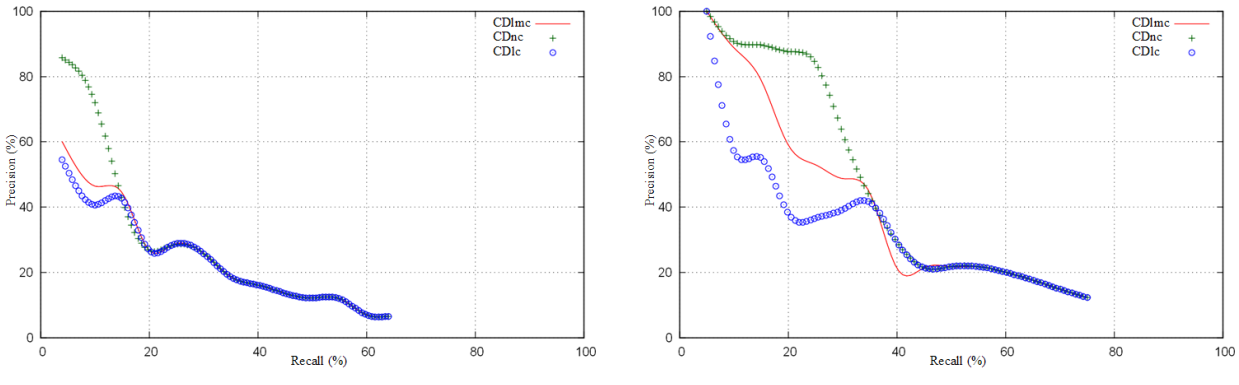


c) d)

Figure 6: Precision-recall curves from Health Education: *a) noun, b) noun-adjective, c) noun-prep-noun, d) all patterns*



a) b)



c) d)

Figure 7: Precision-recall curves from Mathematics: *a) noun, b) noun-adjective, c) noun-prep-noun, d) all patterns*

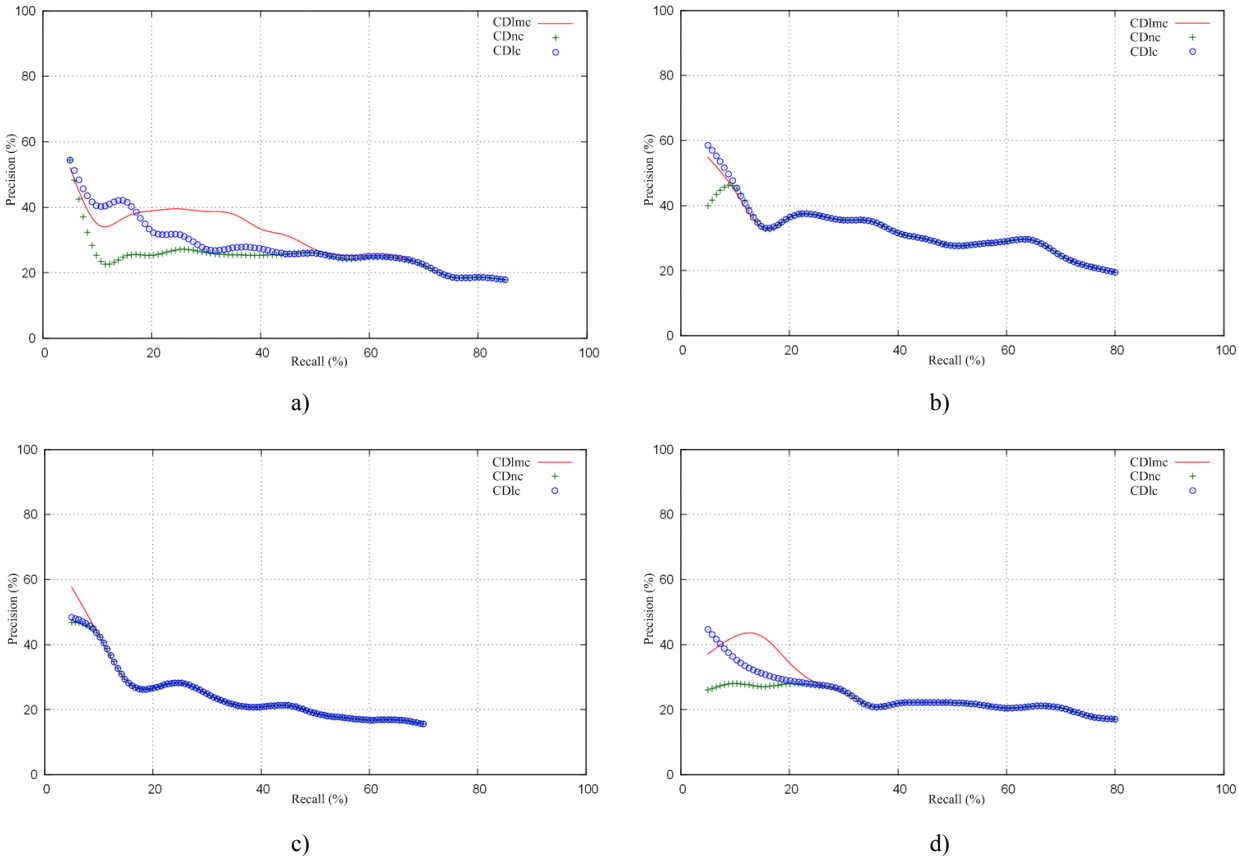


Figure 8: Precision-recall curves from Ecology: *a) noun, b) noun-adjective, c) noun-prep-noun, d) all patterns*

To present the results we first analyse the issues that are common to all domains and then we analyse each domain separately.

The following points may be considered as common to all domains taken in account in this experiment:

- Characteristics of the corpus itself. According to Pearson, (1998), the texts used in the COCIEM may be considered as expert-to-novice communication. These texts usually give good term explanation to ensure they are understood by the reader. Therefore, this set of documents could be considered texts with a low level of specialisation. This situation may create some problems as the use of candidates that may be specialised or not in accordance to the context of use. Consider for example *momento* (instant/momentum) or *multiplicando* (multiplicand/multiplying). Both words may correspond to a specialised concept of mathematics or a word of general use. This fact causes confusion to the extraction tool because it currently looks only for sequences of one or more words without analysing the context or its coincidence with other general uses of such word(s). The actual terminological occurrences of each candidate may require further exploration.
- The lemmatisation of term candidates. A common problem of lemmatisation, especially with the pattern noun-adjective, is the gender agreement. For example, the term *hormonas antidiuréticas* is lemmatised as *hormona antidiurético* instead of *hormona antidiurética*. Thus, term candidates with this type of problem cannot be found in the validation resources and consequently they are tagged as non-related to the domain.
- The Wikipedia itself. The relational database obtained from the snapshot is not perfect; it may happen that following a given link we reach a different page than the expected one. It may also happen that some pages do not have any category registered (ex. *ecuación lineal* -linear equation-, *manada* -herd-, etc). In both cases the term candidate is rejected and wrongly tagged as not belonging to the domain.
- The validated lists. As in any manual task, the human evaluation of a list of term candidates (like ecology and mathematics) is not error-free; thus some actual terms are missing from the validated list causing false errors.

Regarding each of the selected domains there are some considerations to be made.

- Health education

  In this case we used SNOMED-CT, a well known resource in the medicine field. Nevertheless, in spite of its relevance, there are terms, whose termhood is evident, but they are not included in this resource. For example, *medicina* (medicine), the name of the field, is missing. This repository includes more specialised names like *medicina hiperbárica* (hyperbaric medicine), *medicina interna* (internal medicine), *medicina nuclear* (nuclear medicine) among others but not just medicine. Other missing terms are *fisiología* (physiology), *neurotransmisor*

(neurotransmitter) and *alvéolo pulmonar* (pulmonary alveolus). Furthermore, the opposite also holds true, there are candidates like *aceite* (oil), *benceno* (benzene) or *ausencia de enfermedad* (lack of disease) whose termhood may be argued but that are validated as terms because they are included in SNOMED-CT.

- Mathematics
  The list of the terms validated using Wikipedia includes terms that are closely related to the field but do not actually belong to the field. Consider the case of *calculadora de bolsillo* (pocket calculator), its relation to the domain is obvious; however, it does not represent an actual mathematical concept. The reason for this behaviour is the encyclopaedic nature of Wikipedia. This resource includes a redirection mechanism that is frequently used to represent equivalent concepts (like *suma* and *adición* –addition-). Other times it is used as a lemmatization procedure (*matrices* -matrices- redirects to *matriz* -matrix-). And, in Spanish, it is sometimes used to correct very usual typos, e.g. *geometria* redirects to *geometría* (geometry). Due to these typos these cases are not considered terms by the experts but Wikipedia tags them as terms due the redirection system; causing false errors.
- Ecology
  This subject is usually considered a horizontal domain because its concept system embraces terms from several domains like ecology itself but also climate, biology, environment and some others. For this experiment in particular we chose the following categories: Environment, Climate and Biology because we obtain a reasonable result using such set of categories in a similar experiment with English text. It must be considered that not all terms of such areas belong to the field of ecology and the coverage of field by Wikipedia seems to be rather low. These facts are reflected in the results shown in figure 8. Therefore some trade-off is necessary and some error rate must be tolerated. Following this line, clearly terminological sequences like *sumidero de carbono* (carbon sink), *temperatura ambiente* (room temperature) or *eutroficación* (eutrophication) are not considered while others like *enfermedad* (disease), *alvéolo pulmonar* (pulmonary alveolus) or *hipoglucemia* (hypoglycemia) are accepted with the maximum termhood.

4. Conclusions and future work

This experiment allowed us to extract a significant segment of the BSV in Mexican Spanish. In spite of the problems and inconsistencies found in Wikipedia, we have proved that this resource may be useful to validate term candidates in several domains. This has been possible by defining domain borders as a set of Wikipedia categories and calculating domain coefficients in relation to such domain borders. These results are better than those published for similar systems (e.g. Cabré, Estopà & Vivaldi, 2001 and Vivaldi & Rodríguez, 2007). The reason for this may be the lower level of specialisation of the documents. Also, the results are a bit better than those shown in the curves due to the typical issues in term validation.

It is important to note that the proposed tools may be potentially applied to any domain and language providing there is a minimal coverage in Wikipedia.

In the future we plan to experiment with other ways of measuring termhood, such as taking into consideration the number of nodes involved in the paths as well as using the depth of the node in the hierarchy. We also plan to apply this term recognition process to all scientific fields included in COCIEM. This will allow us to compare different educational levels and domains. We still have to establish a better way to determine the categories to be used to validate terms according to the subject and school level. Regarding term candidate lists, we still have to find better ways to validate them (manually or automatically). Another point that still needs to be improved in the treatment of the term candidate list is the relation with longer syntactic patterns; they, usually, as a whole are not terms but a substring of them may be an actual term. Consider the case of *par ordenado de número real* (ordered pair of real number); here the whole expression may be considered a phraseological expression but not a term. Instead, its components (*par ordenado* and *número real*) are terms of this domain. This detection requires further analysis of term candidates (especially longer ones).

As Wikipedia is a continuously updating resource, we plan to improve the results by repeating the experiment for Ecology with newer Wikipedia dumps which may have a higher coverage in this domain.

Our results also show that we need to improve the treatment given to disambiguation pages of Wikipedia, as well as finding a solution for those terms not present as a page but included in the text of articles.

Due to the characteristics of this tool and the kind of texts analysed, it is necessary to build some mechanism for looking at the context in order to distinguish actual terminological usage of the candidates from those that represent general language usage.

Finally we are conscious that the evaluation has been done over the list of term candidates resulting from the extraction processes instead of obtaining them by reading the whole text. This procedure was necessary due to the text size and is not perfect. We plan to choose a subset of the corpus and ask to some specialists to find the terms by actually reading the text and repeat the evaluation procedure.

5. Acknowledgements

This research has received the support from Science and Education Ministry (Spain) for the project "RicoTerm" (HUM2007-65966-C02-01/FILO) and from the Consejo Nacional de Ciencia y Tecnología, CONACyT (Mexico) for the projects 58923 and 82050.

6. References

Arson, A., Lang, F. (2010) An overview of MetaMap: historical perspective and recent advances. *Journal of American Medical Informatics Association*, 17, pp. 229--236.

Barrón-Cedeño, A., Sierra, G., Drouin, P., Ananiadou, S. (2009) An improved automatic term recognition method for Spanish. In: *CICling 2009*. Mexico City: pp. 125--136.

Bourigault, D. (1994) LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes. PhD Thesis. École des Hautes Études en Scienes Sociales.

Cabré, M.T., Estopà, R., Vivaldi, J. (2001) Automatic term detection. A review of current systems. In D. Bourigault, C. Jacquemin and M.C. L'Homme (Eds.), *Recent Advances in Computational Terminology, (*2). Amsterdam: John Benjamins Publishing Company, pp. 53--87.

Cabrera-Diego, L.A., Sierra, G., Vivaldi, J., Pozzi, M. (2011) Using Wikipedia to Validate Term Candidates for the Mexican Basic Scientific Vocabulary. In: *First International Conference on Terminology, Languages, and Content Resources (LaRC 2011)*. Seoul: pp. 76--85.

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

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

Heid, U., Jauß, S., Krüger, K., Hofmann, A. (1996) Term extraction with standard tools for corpus exploration. Experiece from German. In: *Terminology and Knowledge Engineering (TKE'96)*. Berlin.

Kotlerman, L., Avital, Z., Dagan, I., Lotan, A., Weintraub, O. (2011). A Support Tool for Deriving Domain Taxonomies from Wikipedia. In: *Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar: pp. 503--508.

Luyt, B., Kwek, W., Sim, J., York, P. (2007). Evaluating the Comprehensiveness of Wikipedia: The Case of Biochemistry. *Lecture Notes in Computer Science*, 4822, pp. 512--513.

Maynard, D., Ananiadou, S. (2000) Identifying terms by their family and friends. In: *18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken: pp. 530--536.

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

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

Milne, D., Witten, I. H. (2008) An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: *First AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*. Chicago: pp. 25--30.

Panzienza, M.T., Pennacchiotti, M., Zanzotto, F.M. (2005) Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. *Studies in Fuzziness and Sooft Computing*, 185, pp. 225--279.

Pearson, J. (1998) *Terms in context*. Amsterdam: John Benjamins Publishing Company.

Ponzetto, S.P., Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence*, 30, pp. 181--212.

Ponzetto, S.P., Strube, M. (2011). Taxonomy induction based oa a collaboratively built knowledge repository. *Artificial Intelligence*, 175, pp.1737--1756.

Sarmiento, F.O. (2001) *Diccionario de ecología: paisajes, conservación y desarrollo sustentable para Latinoamérica*. Quito: Ediciones Abya-Yala.

Suchanek, F.M., Kasneci, G., Weikum, G. (2008). YAGO: A Large Ontology from Wikipedia and WordNet. *Journal Web Semantics: Science, Services and Agents on the World Wide Web*, 6, pp: 203--217.

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

Vivaldi, J., Rodríguez, H. (2007) Evaluation of terms and term extraction systems: A practical approach. *Terminology*,13(2), pp.225--248.

Vivaldi, J., Rodríguez, H. (2010) Using Wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45, pp. 251--254.

Vivaldi, J., Rodríguez, H. (2011) Extracting terminology from Wikipedia. *Procesamiento del Lenguaje Natural*, 47, pp. 65--73.

Zesch, T., Gurevych, I. (2007) Analysis of the Wikipedia Category Graph for NLP Applications. In: *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*. Rochester, NY: pp. 1--8.

Zesch, T., Müller, C., Gurevych, I. (2008) Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: *The International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech: pp. 1646--1652.