EXPERIMENTS ON TERM EXTRACTION USING NOUN PHRASE SUBCLASSIFICATIONS

Merley S. Conrado
Walter Koza
Universidad de São Paulo
merleyc@icmc.usp.br
kozawalter@opendeusto.es

Josuka Díaz-Labrador
Joseba Abaitua
Solange O. Rezende
Universidad Nacional de Rosario
josuka@deusto.es
joseba.abaitua@deusto.es
solange@icmc.usp.br

Thiago A. S. Pardo
Zulema Solana
Universidad de Deusto
taspardo@icmc.usp.br
zsolana@arnet.com.ar

Abstract

In this paper we describe and compare three approaches for the automatic extraction of medical terms using noun phrases (NPs) previously recognized on medical text corpus in Spanish. In the first approach, as baseline, we extracted all NPs, while for the second and third ones the extraction process is directed to "specific NPs" that are determined on the basis of the syntactic and positional criteria, among others. As contributions (i) we showed that it is possible to extract medical terms using "specific NPs", (ii) new terms were added in the software dictionary, and (iii) terms that were not in the reference lists were extracted. For the third contribution, we used the SNOMED CT® terms lists, aiming at improving the IULA reference lists.

1 Introduction

According to Moreno-Sandoval (2009), generally, noun phrases (NPs) correspond to specific terms of a particular domain. The terms can be formed by only a head or a head and complements. Then, the automatic term extraction task was mainly based on the recognition of this kind of phrases.

In this paper, automatic extraction experiments for medical term extraction using noun phrases (NPs) previously recognized on medical text corpus in Spanish are described and compared. For this task, in a first stage, as baseline, all identified NPs are considered as term candidates, while in the other stages the extraction is directed to "specific NPs" that are determined on the basis of syntactic and positional criteria, among others. The novelty of this work is that we are not using pure noun phrases, like many works utilize. In fact, we are using specific NPs, is to say, a subclassification of phrases. We use the IULA corpus (Bach et al., 1997) of medical texts in Spanish and results are compared with reference lists of unigrams, bigrams and trigrams.

According to the results, (i) we showed that it is possible to extract medical terms using "specific NPs", (ii) the software dictionary was improved with 2,445 new terms, and (iii) other terms that were not in the reference lists were extracted. For the third contribution we used the SNOMED CT® term lists aiming at improving the IULA reference lists. However, it should be mentioned that we detected other expressions that were neither in the reference lists nor in SNOMED CT®, although they could be considered medical terms. In this case, we have to say that new terms are added almost on a daily basis, and it is practically impossible to manually update the terms lists.

2 Term extraction in medicine

There are different works about term extraction that may be applied for different domains, sometimes adaptations are necessary for each of them. For the medical domain, we may mention the contributions of Névéol and Ozdowska (2005) and Bessagnet et al. (2010) for the French; Hao-Min et al. (2008), for the Chinese, and the Lopes et al. (2009), for Portuguese. For the English, we cite the Krauthammer and Nenadic (2004) work, which makes a detailed description of automatic term recognition (ATR) systems in the medical field. Those systems are based either on internal characteristics of specific classes or on external clues that can support the recognition of word sequences that represent specific domain concepts. Different types of features are used, such as orthographic (capital letters, digits, Greek letters) and morphological clues (specific affixes, POS tags), or syntactic information from shallow parsing. Also, different statistical measures are suggested for "promoting" term candidates into terms.

In our work, the term extraction is applied in

the medical domain in Spanish. So here, we mention the main works in this area. We may mention the ONCOTERM Project (Bilingual System of Information and Cancer Resources), the Describe® System, the Vivaldi and Rodríguez works, the Castro et al. works, and the large terminology developed by the SNOMED CT® Project.

ONCOTERM (López Rodríguez et al., 2006) is a Project whose goal is to develop a information system for the oncology domain, in which the concepts are linked to an ontology. The authors worked from Spanish texts to create a terminology database, with correspondences in English and German.

The Describe® system (Sierra et al., 2009), meanwhile, applies a Defining Contexts Extractor (Alarcón, 2009) for the search, classification, and grouping of medical definitions from the web.

Vivaldi and Rodríguez (2010) created a term extraction system that uses Wikipedia (WP) semantic information. It was tested in a medical corpus, and, according to its results, WP was considered a good resource for tasks of medical term extraction.

Castro et al. (2010) work presents a semantic annotation of clinical notes and an application of an automatic tool for medical concept recognition on the SNOMED CT® ontology. Furthermore, a tool test is presented in 100 clinical notes, and, according to the authors, the results are quite good.

SNOMED CT®[1] is a big medical terminology and is the result of the fusion between SNOMED RT and the Clinical Terms Version 3, a terminology previously known as Read Codes, created by the National Health Service (NHS) in England.

3 Term extraction methodology

With the objective of indentifying medical terms, we have developed rules for "specific" NPs recognition. They were used for extracting terms and, as baseline, we consider the term extraction usually performed with NP. We applied it to Spanish, but it may be adapted to others languages, adjusting the linguistic informations of parsers used.

According to Figure 1, the term extraction, carried out this work, starts with the delimitation of the domain and the **corpus**. Afterwards, it is necessary to perform an **orthographic normalization**, changing the corpus file codification to UTF-8. Also, line changes are removed to prevent problems with the tools for the morphological analysis. In the sequence, the **tokenization and morphological analysis** is carried out aiming at tagging words and punctuation marks.

This way, we developed NPs recognition rules (e.g., article + noun = " NP") to shape the NPs to be worked with. Phrase recognition allows the **extraction** of term candidates. At this stage, stopwords are removed of these candidates.

After cleaning the candidates, they are separated into lists of unigrams, bigrams, trigrams and higher than trigrams to allow evaluation.

3.1 Experiments

For the experiments we used the IULA-UPF technical corpus[2] that belongs to the health and medical domains. This corpus is composed of 12 texts in Spanish and the average of words per document is 8,207. With it, the IULA-UPF has also provided three reference term lists, containing a total of 697 unigrams (e.g. "*alergia*" - allergy), 665 bigrams consisting of a name plus an adjective (e.g. "*ácido benzoico*" - benzoic acid) and 82 trigrams formed by a name plus the preposition "de" plus another name (e.g. "*grupo de riesgo*").

From the corpus, we had to recognize noun phrases (NPs), prepositional phrases (PP), and nucleus verbal phrase (nvp).

The term extraction is detailed in Figure 1. The morphological analysis of corpus words was carried out using the SMORPH program (Aït-Mokhtar, 1998), that is a finite-state part of speech tagger that Infosur[3] Group has adapted to Spanish. As an example, for the fragment "*Pruebas de provocación bronquial con ejercicio y con histamina en niños asmáticos.*" (Bronchial provocation tests with exercise and with histamine in asthmatic children.), the test result of SMORPH[4]

[1] SNOMED CT® - `http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html` - "*This material includes SNOMED Clinical Terms® (SNOMED CT®), which is used with permission of the International Health Terminology Standards Development Organisation (IHTSDO). All rights reserved. SNOMED CT® was originally created by The College of American Pathologists. "SNOMED" and "SNOMED CT" are registered trademarks of the IHTSDO.*"

[2] IULA-UPF technical corpus - "*Data belonging to the TECHNICAL CORPUS from Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra (http://bwananet.iula.upf.edu/) in December 2010.*"

[3] Infosur - http://www.infosurrevista.com.ar

[4] References: EMS: morphosyntactic tag; nom: noun; GEN: genre; fem: female; NUM: number; PL: plural; v: verb; ind: indicative; PERS: person; 2a: second, TPO: time; pres: present; TR: type of regularity; irr: irregular; TC: type

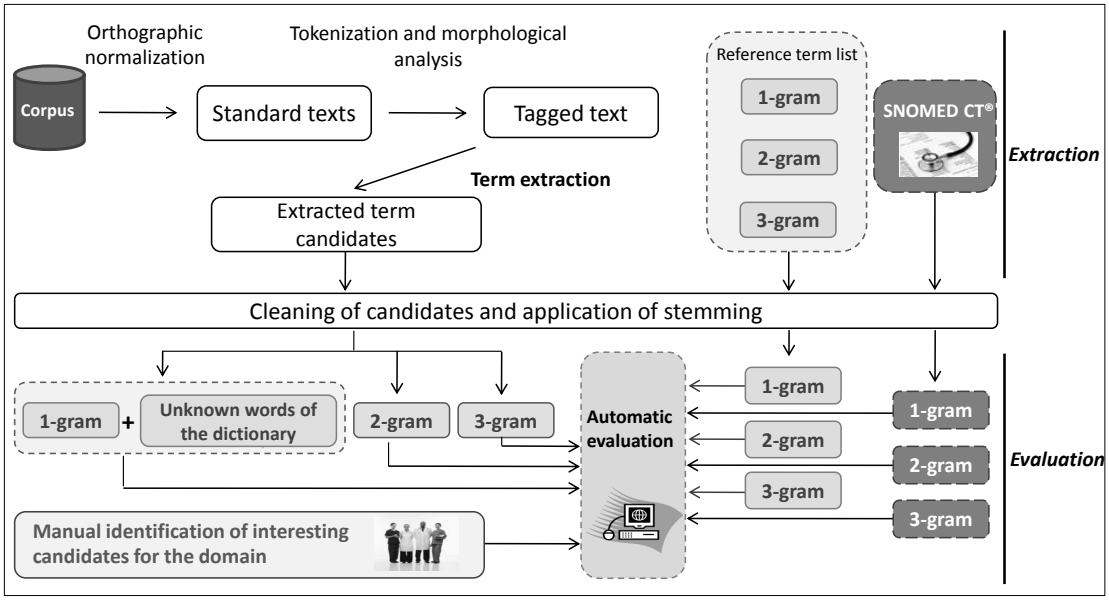

Figure 1: Term extraction and evaluation methodology.

is showed in Table 1. A total of 2,445 words of this corpus were not identified by the parser. This way, they were manually analyzed and added to the original dictionary of the program.

'Pruebas'.
['prueba', 'EMS','nom', 'GEN','fem', 'NUM','pl'].
['probar', 'EMS','v', 'EMS','ind', 'PERS','2a', 'NUM', 'sg', 'TPO','pres', 'TR','irr', 'TC','c1', 'TDIAL','est'].
'**de**'. ['de', 'EMS','prde'].
'provocación'.
['provocación', 'EMS','nom', 'GEN','fem', 'NUM','sg'].
'bronquial'.
['bronquial', 'EMS','adj', 'GEN','-', 'NUM','sg'].
'**con**'. ['con', 'EMS','prep'].
'ejercicio'.
['ejercicio', 'EMS','nom', 'GEN','masc', 'NUM','sg'].
'**y**'. ['y', 'EMS','cop'].
'**con**'. ['con', 'EMS','prep'].
'histamina'.
['histamina', 'EMS','nom', 'GEN','fem', 'NUM','sg'].
'**en**'. ['en', 'EMS','prep'].
'niños'.
['niño', 'EMS','nom', 'GEN','masc', 'NUM','pl'].
'asmáticos'.
['asmático', 'EMS','adj', 'GEN','masc', 'NUM','pl'].
'**.**' ['linsig', 'EMS','pun'].

Table 1: Morphological analysis SMORPH.

In the sequence, noun phrase recognition rules were developed. These rules are loaded into the MPS syntactic parser (Abbaci, 1999) that receives the SMORPH output as input.

Three different experiments were performed considering the noun phrase sub-classification.

For the first experiment (**Exp. NP**), all expressions previously tagged as NPs were considered as term candidates. For the second one (**Exp. S_NP**), after manual observations about the terms, some NP that could be relevant were subclassified. This subclassification considered the possibility that:

- the NP could be a verbal argument (NP_VARG): "*detectó la bronconeumonía*" (He *detects bronchopneumonia*). For it, the rule corresponding to the structure *NP + svn = NP_VARG* was created.

- the NP could be an antecedent of a non-defining clause (NP_NONDEF): "*el asma, que se traduce...*" (asthma, which means). Here we took several rules and an example of them is *NP + coma + relative + svn = NP_NONDEF*. Rules for non-defining clause recognition were created. For this work, we only considered that expression from the NP-antecedent until verb clause.

- the NP could be an item from an enumeration (NP_ENUM): "*dolor de garganta, fiebre y tos*" (headache, fever, and cough). An example of enumeration rule is *NP + coma + NP + conjunction + NP = NOM_COMP_ENUM* (Nominal complete enumeration).

- the NP could be in parentheses (NP_PARENT): *(fenoterol)*. The rule corresponding to the structure *parentheses*

of conjugation; c1: first conjugation; TDIAL: type of dialectal variety; est: standard; prde: preposition "*de*"; prep: preposition; masc: male; cop: copulative; sg: singular, linsing: next line; pun: dot.

+ NP + parentheses = NP_PARENT was created.

- the NP could be at the beginning of the clause (NP_INIC): "*...en los últimos años. El mecanismo inmunológico es...*" (...in recent years. The immunological mechanism is...). In this case, for the construction of the rule, the endpoint of the previous sentence was considered: *endpoint + NP = NP_INIC*. NP that appears at the beginning of clause was regarded as a candidate, because the candidate of this sentence position could be the subject or it could be a topicalized element. This rule considered that subjects and topicalized elements are relevant to the terminology extraction.

- the NP could be a argument of a prepositional phrase (PP) at the beginning of the clause (NP_PPINIC): "*...infección bacteriana. Para el diagnóstico...*" (...bacterial infection. For diagnosis...). In the same way as in the previous case, the endpoint of the sentence was considered: endpoint + preposition + NP = NP_PPINIC.

In the third experiment (**Exp. S_NP2**), we used the subclassification of *Exp. S_NP* and the NPs that are PP arguments were added: "*en estudios epidemiológicos*" (in epidemiological studies).

In all experiments, the **cleaning** of the extracted terms was carried out aiming at removing the numerals. This cleaning consists of discarding of candidates composed only of one letter, stopwords from the extremities of the candidates, and candidates that fully corresponded to stopwords. We used the stoplist available in the Snowball Project[5] and we added verb conjugations *poder* and *deber* and some words such as *año* (year), *días* (days), *algún* (any), etc., totaling 733 stopwords.

Also, in the case of NP_VERB, the right extremities *svn* were removed. For example, in the NP_VERB "*se detectan 636 asmáticos*" - (636 asthmatics were detected), after removing "se detectan" and cleaning this example, the candidate was reduced to: "*asmáticos*" (asthmatics).

Subsequently, in order to allow further evaluation, term candidates were separated into term lists of unigrams, bigrams, trigrams.

[5]Snowball Project - http://snowball.tartarus.org/algorithms/spanish/stop.txt

3.2 Results and evaluation of experiments

The number of extracted candidates is showed in Table 2.

	Unigrams	Bigrams	Trigrams
Experiment NP	1744	2684	1999
Experiment S_NP	856	1172	824
Experiment S_NP2	1188	1913	1419

Table 2: Number of extracted candidates.

Two automatic tests were carried out (Figure 1). In the first one, IULA **reference lists** were used to verify the quality of extracted candidates.

First of all, it was necessary to apply **stemming** techniques (PreTexT II tool (Soares et al., 2008)) to the extracted terms and reference term list, due to morphological variations in the words. Subsequently, it was possible to compare the extracted terms and the reference term list.

The accuracy and coverage for all three experiments (NP, S_NP and S_NP2) are showed in Figures 2, 3, and 4, respectively, for unigrams, bigrams, and trigrams. The figures are modified from Vivaldi and Rodríguez (2010) because they used the same corpus in their experiments, so, we also present a comparison between our and their results. In their work, *EWN* corresponds to the group of extracted terms using the YATE method (Vivaldi, 2001). The other terms were extracted with the Wikipedia categories (WP) having "Medicina" as domain name and varying the calculation of the domain coefficient. In *WP.lc*, the number of simple steps given in Wikipedia is considered; *WP.lmc* takes into consideration the mean number of paths in Wikipedia; *WP.nc* takes into consideration the number of paths in Wikipedia. It is important to notice that the extraction proposal of Vivaldi and Rodríguez only considered patterns with the following structures: (i) *noun* (for unigrams), (ii) *noun + adjective* (for bigrams), and (iii) *noun + the "de" preposition + noun* (for trigrams). This highly contrasts with our extraction that considers all possible combinations.

For the second test, the quality of the candidates was verified according to the SNOMED CT® list, which has 1,060,632 Spanish terms. Subsequently, the candidates that could be interesting for the medical domain were manually identified and, afterwards, we checked if those candidates were present or not in the SNOMED CT® list. The verification was done separately for each experiment (Exp. NP, Exp. S_NP, and Exp. S_NP2) and the results were separated into unigrams, bi-

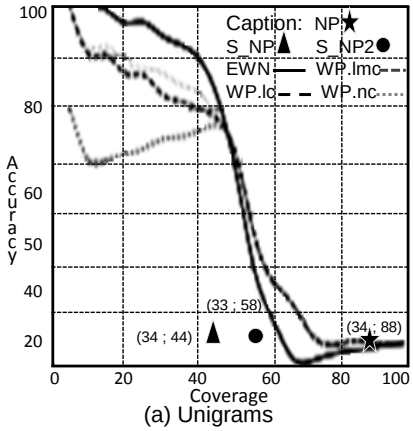

Figure 2: Accuracy and coverage values obtained for unigrams.

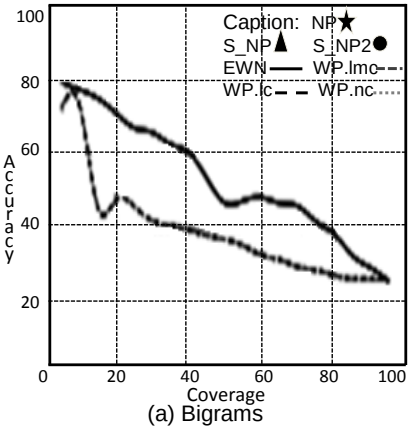

Figure 3: Accuracy and coverage values obtained for bigrams.

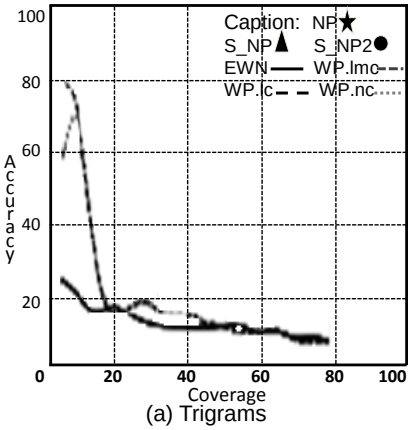

Figure 4: Accuracy and coverage values obtained for trigrams.

grams, and trigrams. The candidates that could represent terms according to the SNOMED CT® list are showed in Figure 5.

It is quite difficult to get a constant and immedi-

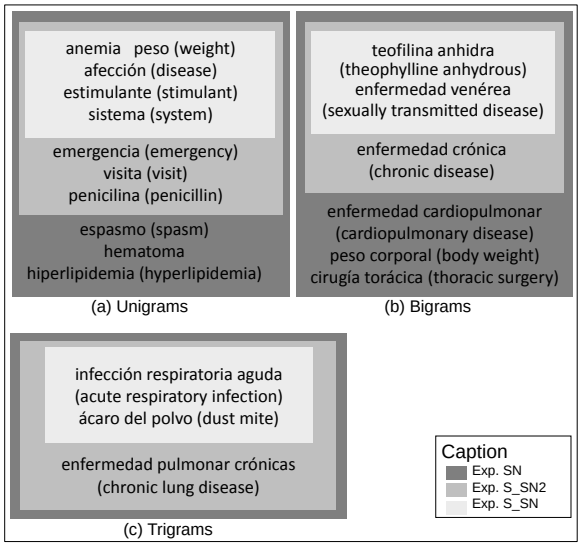

Figure 5: Extra terms obtained.

ate updating on medical terminology (Krauthammer and Nenadic, 2004). This fact motivated us to perform a **manual identification of candidates that are interesting for the medical domain**. These candidates were not present in the reference lists nor in SNOMED CT®, although they seem to be important for this specific domain. Here we present some examples: "*insuficiencia ventilatoria obstructiva*" (obstructive ventilatory failure), "*paciente asmático atópico*" (atopic asthmatic patient), (respiratory atopic diseases), "*traumatismo encéfalo craneano*" (traumatic brain injury), etc.

4 Conclusions

If we compare the three experiments carried out (NP, S_NP, and S_NP2), little accuracy variations are found for unigrams, bigrams, and trigrams, although the coverage varies in each case. We were able to obtain the best coverage in the first experiments, in which we took all NPs as term candidates. Nevertheless, we expected those results because most of the candidates are obtained when all NPs are extracted, and it allows for a large coverage. However, we expected better accuracy rates for the cases with "specific NPs".

In the comparison, we may see that the results obtained were similar to those of Vivaldi and Rodríguez in the case of unigrams, although they were able to obtain better results for bigrams and trigrams. Regarding this fact, we observed that the best accuracy rate was achieved with the experiments in which the NPs were part of an enumeration. Also, we emphasize the simplicity of our ex-

traction method, which does not require external knowledge and was able to work well using the SMORPH dictionary and MPS recognition rules, also not considering only reference list patterns but all possibilities. In addition, better accuracy is expected by new and more specific MPS rules.

According to the results, we obtained three interesting contributions: (i) we were able to show the possibility of extracting medical terms from recognition of "specific NPs", even that it is necessary improvements in the method; (ii) the SMORPH dictionary was improved with 2,445 new terms. Thus, we expect to have better experiments in the medical domain with this tool; (iii) other terms that were not present in the reference lists were also extracted. Those terms were tested with the SNOMED CT® and we obtained terms that could be added to the IULA reference lists, which means an improvement of these lists. At the same time, we observed that there were other terms with a different structure from "noun + the 'de' preposition + noun". This evidences the fact that there exists important trigrams that do not necessarily fit to that pattern.

As future work, we intend to improvethe accuracy with new filtering rules, to increase the SMORPH dictionary, and to test the extraction rules in larger corpora and other domains.

Acknowledgments

Thanks to Erasmus Mundus, CNPq, FAPESP y CONICET for financial support and to Vivaldi y Rodríguez for making available the dataset.

References

F Abbaci. 1999. Développment du module post-smorph. In *Memória del DEA de Linguistique et Informatique*. Groupe de Recherche dans les Industries de la Langue - Universidad Blaise-Pascal - Clermont-Ferrand.

Rodrigo Alarcón. 2009. *Extracción automática de contextos definitorios en corpus especializados*. Ph.D. thesis, Universidad Pompeu Fabra, Barcelona.

S Aït-Mokhtar. 1998. *L'analyse présintaxique en une seule étape*. Ph.D. thesis, Groupe de Recherche dans les Industries de la Langue - Universidad Blaise-Pascal - Clermont-Ferrand.

Carme Bach, Roser Saurí Colomer, Jordi Vivaldi, and M. Teresa Cabré Castellví. 1997. El corpus de l'IULA: descripció. Technical Report 17, Universitat Pompeu Fabra – Institut Universitari de Lingüística Aplicada, Barcelona - Spain.

Marie-Noëlle Bessagnet, Eric Kergosien, and Mauro Gaio. 2010. Extraction de termes, reconnaissance et labellisation de relations dans un thésaurus. *CoRR*, abs/1002.0215.

Elena Castro, Ana Iglesias, Paloma Martínez, and Leonardo Castaño. 2010. Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 751–757, New York, NY, USA. ACM.

Li Hao-Min, Ying Li, Hui-Long Duan, and Xu-Dong Lv. 2008. Term extraction and negation detection method in chinese clinical document. *Chinese Journal of Biomedical Engineering*, 27(5).

Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *J. of Biomedical Informatics*, 37:512–526, December.

Lucelene Lopes, Renata Vieira, Maria Finatto, Daniel Martins, Adriano Zanette, and Luiz Ribeiro Jr. 2009. Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde - doi: 10.3395/reciis.v3i1.244pt. *Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, 3(1).

Clara Inés López Rodríguez, Maribel Tercedor, and Pamela Faber. 2006. Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. *Revista E Salud*, 2(8).

A. Moreno-Sandoval. 2009. Terminología y Sociedad del conocimiento. pages 99–116. Peter Lang.

Aurélie Névéol and Sylwia Ozdowska. 2005. Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français. In *EGC*, pages 655–666.

Gerardo Sierra, Rodrigo Alarcon, Alejandro Molina, and Edwin Aldana. 2009. Web exploitation for Definition extraction. In *Proceedings of the 2009 Latin American Web Congress*, pages 217–223, Washington, DC, USA. IEEE Computer Society.

M. V. B. Soares, R. C. Prati, and M. C. Monard. 2008. Pretext II: Descrição da reestruturacão da ferramenta de pré-processamento de textos. Technical Report 333, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP - São Carlos, São Carlos - SP.

Jorge Vivaldi and Horacio Rodríguez. 2010. Using wikipedia for term extraction in the biomedical domain: first experiences. *Procesamiento del Lenguaje Natural*, 45:251–254.

Jorge Vivaldi. 2001. *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Ph.D. thesis, Universitat Politécnica de Catalunya, Barcelona, Spain.