PROPOSAL FOR AUTOMATIC EXTRACTION OF MEDICAL TERM CANDIDATES WITH LINGUISTIC INFORMATION PROCESSING DESCRIPTION AND EVALUATION OF RESULTS

Walter KOZA ORELLANA[*]

- ABSTRACT: The description of a method for automatic extraction of term candidates from the medical field by applying linguistic information is presented. Lexicography, morphological and syntactic rules were used. First, the detection was performed by applying a standard dictionary that assigned the tag ´MED´ (‘MEDICAL’) to the words that could be considered terms. Morphological and syntactic rules were used to try to deduce the part of speech of the words that were not considered in the dictionary (WNCD). Afterwards, nominal phrases that included WNCD and MED were gathered to extract them as term candidates of the field. Smorph, Post Smorph Module (MPS) – both work in groups– and Xfst were the software used. Smorph performs the morphological analysis of character strings and MPS works on local grammar. Xfst is a finite state tool that works on character strings assigning previously stated categories to allow the automatic analysis of expressions. This method was tested on a section of the corpus of clinical cases collected by Burdiles (2012) of 217258 words. The results showed 92.58% of precision, 95.02% of recall and 93.78% of F-measure.

- KEYWORDS: Medical terminology. Automatic extraction. Linguistic information. Terms candidate.

Introduction

The unprecedented development of communication technologies has enabled, mainly from the Internet, the production, access and exchange of a huge flow of information and scientific knowledge to people around the world. However, to access this great amount of data, it is necessary to have tools that can process data, with systems of storage and retrieval of information (LÓPEZ-HUERTAS; BARITÉ; TORRES, 2004). At the same time, it is also essential to develop resources to regulate and discuss the concepts of the various areas of knowledge, as well as assigning new names for new concepts, with the aim of ensuring an adequate scientific communicability. Studies on the area of computational linguistics have made several contributions to information retrieval systems (VILLAYANDRE,

[*] PUCV – Pontificia Universidad Católica de Valparaíso. Instituto de Literatura y Ciencias del Lenguaje. Facultad de Filosofía y Educación. Viña del Mar – Valparaíso – Chile. 2530388 – walter.koza@ucv.cl

http://dx.doi.org/10.1590/1981-5794-1502-5

2010) allowing users to access data faster and more accurately. One of the main tasks in the development of these systems is the automatic detection of domain-specific terms. A term is a lexical unit that represents a concept in a particular subject field (SAGER 2000; MARINCOVICH, 2008). From a corpus linguistics point of view, the output of a terminology process can also be considered as a term (JACQUEMIN; BORIGAULT, 2005). The extraction of terms representing an area usually constitutes the starting point for more complex tasks, such as making lists of entries for specialized dictionaries, creating databases or ontologies and taxonomies that organize and specify the area of knowledge, etc. The main disadvantage is the constant change of terminology, which hinders the manual update of terminology databases and implies the need for tools that can detect new terms and their variations (KRAUTHAMMER; NENADIĆ, 2004). Moreover, the extraction tasks, especially those that appeal to linguistic analysis techniques, tend to focus on specific areas of knowledge, in order to adapt to the requirements and characteristics of each one in particular.

Now, one of the fundamental areas of knowledge is medicine, not only for its social role, preserving the physical integrity of human beings, but also for the increasing production and circulation of data related to this area (articles, case reports, reports, etc.). To this end, the method developed for extracting term candidates of the medical field from linguistic information processing is described in this paper. This work is framed in the field of computer language, on one hand, and text mining tasks, on the other.

According to Cabré (2006), the complexity of automatic term detection lies in developing a processor with the same abilities of a human specialist; this point of view could be considered as extreme given that it would be impossible to create and extractor with such skills. However, it is possible that machines process the same information as a human specialist. This information is semantic, morphological and syntactic. For this purpose, the rules given for the developed method were based on the above mentioned information and tested on a section of the corpus of clinical cases collected by Burdiles (2012).

At a lexical level, detection was achieved with the use of a standard dictionary, in this case, the *Diccionario Esencial de la Lengua Española* (Essential Dictionary of the Spanish Language) (DICCIONARIO…, 2006). The dictionary was uploaded to the analysis software which assigned the tag 'MED' ('MEDICAL') to the words considered as terms; for this task, experts of the field identified those words from the RAE dictionary that belonged to the medical field. The following assumption was established for the rest of the words: the words that were not considered in the dictionary (WNCD) and can be classified as noun or adjective are, mostly, specific expressions of the medical field. Worth mentioning that, in this study, we took into account the proposition of Moreno-Sandoval (2009), which provides that, generally, noun phrases correspond to terms. To this end, the extraction tasks were focused on those noun phrases.

Word formation and syntactic rules were used to try to deduce the part of speech of the words that were not considered in the dictionary (WNCD). Afterwards, noun phrases that included WNCD and MED gathered to extract them as term candidates of the field. Finally, the method's precision, recall and F measure were assessed.

The computer processing was done with Smorph (AÏT MOKTHAR, 1998), Post Smorph Module (MPS) (ABACCI, 1999) and Xfst (BEESLEY; KARTTUNEN, 2003). The first one performs the morphological analysis of the character string, which yields morphological and POS allocation for each occurrence according to the features given. MPS, in turn, uses the output of Smorph as its input and, from regrouping, ungrouping and correspondence rules established by the user, analyzes the headword string that results through the morphological analysis. Xfst is a finite state tool that works on character strings assigning previously stated categories to allow the automatic analysis of expressions. This requires a set of rules that interact to establish possible combinations of categories.

The paper is organized as follows: Section 2, previous works on the area, Section 3, methodology and work done, Section 4, results, and Section 5, conclusions of the research.

Term extraction in the medical field

Regarding the medical field, Krauthamer and Nenadić (2004) state that the conditions for a successful term extraction include lexical variations, synonymy and homonymy. On the other hand, it is difficult to keep the terminological resources updated due to the constant change of terminology. Some terms are used for short periods of time and new terms are included to the vocabulary of the field almost every day. Furthermore, we must add the lack of strict conventions for nomenclatures. There are guidelines for some types of medical entities, but these guidelines do not set any limitation to the experts; therefore, they are in no way forced to use such guidelines when establishing a new term. Also, along with 'well-formed' terms there are *ad hoc* names, which are problematic for term identification systems. However, despite the difficulties mentioned, various systems for detection of terms have been developed for many kinds of medical institutions. These systems are based on internal features of specific classes and on external cues that might help to identify strings of words that represent concepts of the domain. Different types of features are used, such as spelling (upper case, digits, and Greek characters) and morphological cues (specific affixes and formants) or information from syntactic analysis. In addition, different statistic measures are suggested to consider term candidates as terms.

For Spanish, the works done by López, Tercedor and Faber (2006) for Oncoterm project are worth mentioning. The project is an interdisciplinary research about terminology with the aim of developing an information system for a medical subarea, oncology, where concepts are linked to ontology. To do this, they use information from specialized dictionaries and corpus, as well as dictionaries and corpus provided by experts.

Castro et al. (2010), meanwhile, presented a proposal for detection of concepts of clinical notes, implementing a tool to identify biomedical concepts in SNOMED CT (IHTSDO, 2013). They describe the process of semantic annotation of terms in the ontology on a corpus of clinical notes. The experiments focused on comparing the automatic labeling of SNOMED CT with the manual annotation carried out by experts in the field. According to the authors, the functionalities of the tool let you obtain more semantic knowledge, affecting the establishment of new relationships that allow text mining in clinical notes.

In turn, based on SNOMED CT and other ontologies like UMLS (NLM, 2013) there have been studies of automatic recognition of semantic similarity. Among them, there can be mentioned those carried out by Sanchez, Batet and Valls (2010), and Garla and Brandt (2012). Both works are focused on analyzing automatically the relationship between concepts that share the same context.

On the other hand, using semantic information extracted from Wikipedia, Vivaldi and Rodriguez (2010) present a term extraction system tested on a medical corpus. The experiments consist of taking a document and a corresponding set of term candidates and compare the results obtained using EuroWordNet and Wikipedia. This involves exploring the second resource in order to obtain a domain coefficient equivalent to that obtained with EWN. This method has the following steps, for a given term candidate: (i) find a corresponding Wikipedia page, (ii) find all categories of Wikipedia associated with that page, and finally (iii) examine Wikipedia accessing recursively to all the links of the categories found in (ii) to enrich the domain edge. According to the authors, the results show that this resource can be used for tasks of automatic term extraction.

Finally, on the field of translation and corpus linguistics, Moreno-Sandoval and Campillo-Llanos (2013) develop a corpus of biomedical texts in Spanish, Arabic and Japanese. The texts included in this corpus are not extremely technical, but targeted at medical students, for example, manuals and medical journals for the public in general. The purpose of the authors is to develop a term search engine with that corpus for three languages and compare them.

With regards to the approaches based on linguistic knowledge can be divided in two of them, those based on dictionaries and those based on morphological and syntactic rules. Methods based on dictionaries use existing terminology resources for the purpose of locating the occurrences of words in texts. The evident limitation of these methods is that many occurrences cannot be recognized by standard dictionaries or standard databases, however, in this study, it can be seen that having lexicographical information of dictionaries provides an ideal base for term extraction tasks. Also, homonym and different spellings of one term can have a negative effect, for example, different with the use of punctuation marks (bmp-*4 / bmp4*), different numerals (*syt4 / sytiv*), different transcriptions of Greek letters (*igα / igalpha*) or different order (*integrin alpha 4 / integrin4 alpha*) (TUASON et al., 2004).

On the other hand, methods based on rules, in turn, try to retrieve terms using the same composition patterns used to build the terms in natural language. The main issue with these methods is to develop rules describing common naming structures for certain types of terms using orthographic or lexical cues, as well as, more complex morphosyntactic features. From this perspective, we can mention the work of Segura, Martinez and Sami (2008), focused on automatic detection of generic drugs using the metathesaurus ULMS and naming rules to create generic drugs proposed by the board of United States Adopted Names (USAN) (AMA, 2013), which allows the classification of drugs in drug families. With this technique one can detect drugs not included in UMLS. The authors achieved 100% coverage and 97% accuracy using UMLS, and 99.3% precision and 99.8% coverage using a combination of lexical information given by UMLS and rules of formation of drug names proposed by USAN. Subsequently, Gálvez (2012) proposed a similar work but based solely on morphological rules, like Segura, Martinez and Sami (2008), proposed by USAN and using the finite state tool NooJ (2013). Thus, the author achieves 99.8% accuracy and 92% coverage.

The method presented here uses two approaches, i.e. information provided by dictionaries, in this case, we chose a standard dictionary, and deduction of words not included in this dictionary using morphological cues. Furthermore, information provided by syntactic context is also used. On the next point, the work done is described.

Machine modeling and implementation

The elaboration of a set of semantic, morphological and syntactic rules for the detection of appropriate terms of the medical field was carried out in order to develop an automatic-detection method of term candidates of the mentioned area.

The procedure of this work is based on two fundamental aspects: (i) the assignment of the semantic tag 'med' (which stands for 'medical') to the entries of the Smorph dictionary in order to recognize, in the texts, those terms specific to the medical field that can be found in a standard dictionary, this task was applied only for unigram detection; and (ii) the deduction of part of speech of words that cannot be found in the source dictionary of Smorph through: (a) its morphological structure and (b) its syntactic context. The terms of the area that can be found on the Essential Dictionary of the Spanish Language (DICCIONARIO…, 2006) (for example: 'enfermedad', 'médico', 'cáncer', 'presión baja', among others) were compared for the first aspect. On point (a), experts of the field helped to identify entries from the RAE dictionary that belonged to the medical field. On point (b), studies on overall word formation (VARELA, 2005) and medical word formation (DURUSSEL, 2006) were considered; as well as the relationship of morphology and terminology (CABRÉ, 2006) and the analysis of shaping phrases (NUEVA…, 2010).

For the computational work, Smorph (AÏT MOKTHAR, 1998), Post Smorph Module (MPS) (ABACCI, 1999), and Xerox's Xfst (BEESLEY; KARTTUNEN, 2003) were used.

Smorph is an analyzer and text generator. On a single step, it isolates and analyzes, morphologically, the text segments to consider, shaping entries with their corresponding values. This software is a declarative tool and the data used is apart from the algorithmic machine, this means it can be adapted to the user's needs. The same software can handle any language as long as the linguistic information is changed.

Smorph declarative sources consist of five files: (i) ascii.txt: it contains the specific ascii codes, such as sentence and paragraph splitter; (ii) rasgos.txt: it includes labels of morphological features that are applied in the analysis of character strings with its possible values (for example, EMS: 'name', 'verb'; gender: 'masculine', 'female', among others); (iii) term.txt: it loads the different endings (similar to suffixes but not the same) that each headword may present in its morphological derivation (e.g.: -o -a, -os, -as); (iv) entradas.txt: it is the list of corresponding headwords and models of derivation (e.g.: casar v1); and (v) modelos.txt: it defines the classes according to the parameters of concatenation of strings from entries and endings (e.g.: Model v1: root word + endings of the 1° regular conjugation + features). One of the features of the program is that default categories can be allocated. In this case, the label 'UW' ('unknown word') is automatically allocated to those words that are not part of its dictionary. At the same time, it can also classify words in relation to its ending, which Aït Mokthar (1998) refers to as "distinguished endings", for example, all Spanish words finished with "-ción" are female nouns; for this reason, loading nouns with that ending will not be necessary, since it will be enough to indicate that information in the term.txt file.

On the other hand, the MPS declarative sources are formed by a unique type of file: rcm.txt, which includes a list of rules that specify possible headword strings with a computerized syntax. There are three types of rules:

1. Regrouping: Determinant + Noun = Noun Phrase
2. Ungrouping: Contraction = Preposition + Determinant
3. Correspondence: Article = Determinant

Lastly, in the case of Xsft, the application is presented as an implementation of finite state machines. Its aim is to produce a morphological analysis and generation. This tool works with source files in which linguistic information is declared in a plain text editor (.txt). Some of the tools that are used by the program are the finite state tokenizers that run the segmentation of the text according to the stored morpho-syntactic information. In this case, this tool was used in order to identify medical terms that are formed by any typical medical formant, for instance: '*-algia*', for '*neuralgia*', *'gastralgia'*; '*blasto-*', for '*blastocito*', '*blastoma*', and so on. The UW recognition process and the subsequent term candidates' extraction have the following stages:

Stage I: Morphological analysis and recognition of punctuation marks using Smorph. In this step, the label 'UW' was assigned to the unknown words.

Stage II: Modification of the term.txt file through the assignment of distinguished endings with its corresponding morphological classification. Subsequently, the corpus is run by Smorph again in order to obtain the categories that can be adjusted to those endings. Additionally, in this stage, it was possible that the UW was a proper noun or an abbreviation, depending on whether capital letters were considered or not.

Stage III: Recognition of the term candidates from morpho-syntactic structures through Xfst. The corpus was run by Xfst with the purpose of detecting those words that contain in its structure any distinctive feature of a medical term. For this reason, as an example, rules of the following kind were stated in the source file: 'necro + letter(s) = medical term' (examples: '*necropsia*', '*necrosis*'); 'letter(s) + cardio + letter(s) = medical term' (examples: '*microcardiopatía*', '*electrocardiograma*'). Those words recognized by this method were labeled 'UW' and were adjusted to the output format of Smorph.

Stage IV: Creation and implementation of syntactic rules that allow the deduction of PD categories. Here, the noun phrase (SN – 'sintagma nominal') is emphasized, (e.g.: Det + PD + Adj = SN/ART + NOM + ADJ).

Stage V: Extraction of SN that involve PD, as term candidates. The terms were simplified using the stemming technique (MANNING; RAGHAVAN; SCHÜTZE, 2009). This technique reduces words to its non-inflectional and non-derivative forms.

Stage VI: Assessment of the categorizations and the term candidates extracted under expert guidance.

The suggested method was tested on a section of the corpus of clinical cases, CCCM-2009, collected by Burdiles (2012). This corpus includes clinical cases covered in medical journals. A brief extract of the corpus is used as an example, where a set of specific terms were recognized.

Figure 1 – Extract taken from the analyzed CCCM-2009

> Enfermedad de tricocefalosis es la infección por **Trichuris trichiura**, parásito que se ubica en el intestino grueso, que con frecuencia se comporta como comensal, pero puede originar sintomatología cuando está presente en gran número, especialmente en niños con deficiencias nutritivas. (Boletín Chileno de Parasitología, v.54, n.3-4, 1999).

Fonte: apud Burdiles (2012).

Smorph tagged '*enfermedad*', *'infección*', '*parásito*', '*intestino*', *'comensal'*, '*sintomatología'* and '*desnutrición'* with TC tag, since they were part of the source dictionary. At the same time, '*tricocefalosis*', '*Trichuris*' and '*trichuria*' were tagged with UW tag. These words were identified through the aforementioned stages.

1. The text was analyzed by Xfst, in which the file with the rules of morphological level had:

   letter ≥ 1 + cefal + letter ≥ 1 = 'CT' (4)

   It is important to clarify that the expression 'cefal' is part of the list of medical roots.
2. Then, it was analyzed by MPS, where the syntactic rules file, rcm.txt, included:

   Preposition + UW + UW + Punctuation Mark = Prep_SNMED_SigP (5)

   CT + preposition 'de' + CT = Trigram (6)

For the expressions tagged as Prep_MEDNP_PM ('Preposition_Medical Noun Phrase_Punctuation Mark'), the preposition and the punctuation mark obtained in the bigram 'Trichuris trichuria' were deleted.

The suggested method was assessed through precision and recall measures. The results will be shown in the next section.

Evaluation

The results of the experiments were evaluated according to accuracy, coverage and f measures. The experts of the field made a reference list of 10092 terms divided as follows:

- Unigramas: 2367
- Bigrams: 5084
- Trigrams: 2641

From the 10092 term list, 9590 were correctly recognized and 769 were wrongly classified. This translates to 92.58% accuracy, 95.02% coverage and 93.78% f measure. The table below shows the results divided in unigrams, bigrams and trigrams.

Table 1 – Results

	Unigrams:	**Bigrams:**	**Trigrams:**
Accuracy	79.65%	96.96%	99.25%
Coverage	97.08%	91.48%	96.02%
F Measure	87.50%	94.14%	97.61%

Source: Made by the author.

As can be seen, the best accuracy was obtained for trigrams, while the best coverage was achieved for unigrams; also, the f measure was best for trigrams.

Some issues with precision for unigrams were detected; one of the causes was that some common words had some elements in common with the terms, such as '*fotografía*'. In terms of coverage, issues were caused by medical words not being considered as such in the RAE dictionary, for example, '*diámetro*'. Also, several spelling errors by the authors affected the results.

However, from the results obtained, this can be considered as a valid method.

Conclusions and further work

An automatic-detection method of term candidates of the medical field through the application of linguistic techniques was presented. For this purpose, we worked with rules at the semantic, morphological and syntactic level using Smorph, Post Smorph Module (MPS) and Xfst software.

The proposed method was tested in a subset of the corpus of clinical cases CCCM-2009 collected by Burdiles (2012), achieving 95.02% coverage, 92.58% accuracy and 93.78% f measure. The obtained results suggest that this method is, roughly, effective and opens up new perspectives about the automatic extraction of term candidates.

It is important to note that a standard dictionary to test the effectiveness of morphological and syntactic rules was chosen.

From the results, it was observed that approximately 50% of the terms not found in the Essential Dictionary of the Spanish Language (DICCIONARIO…, 2006) were detected by these rules.

Nevertheless, in future experiments, the work will be done with a dictionary of the field, *Diccionario de términos medicos* (2012) (Dictionary of medical terms), of the Real Academia de Medicina and the results will be compared.

The errors in term detection were mainly caused by UW with morphological structure different from the one medical terms have and were, instead, isolated or the fact that the surrounding elements were not enough to deduce its part of speech, for example a vertical list or a list in parenthesis. It is important to mention that the cases of proper nouns that, in some occasions, can be terms as is the case of "Alzheimer", implied that they cannot be rejected in the first place. Lastly, the amount of spelling and writing errors of some texts made a negative impact on the results.

The main advantage of this type of method is that its effectiveness can be demonstrated not only for a great amount of texts, but also in smaller corpus, with fewer words. This should help in automatic classification tasks of documents from the extracted terms.

This paper aims to contribute to the task of data extraction, as well as for studies of medical terminology, introducing the analysis of morphological structure of texts and studying the syntactic contexts in which such constructions appear.

Future work is organized around the following axes:

First, is to add specific lexical information of the *Diccionario de términos médicos* (2012). Second, is to try to add the statistical techniques to the proposed method. Third, analyze and develop rules for the automatic detection of word variation. Finally, fourth, consider possible techniques for automatic classification of documents from the terms extracted by this method.

KOZA ORELLANA, W. de. Propuesta de extracción automática de candidatos a término del dominio médico procesando información linguística: descripción y evaluación de resultados. **Alfa**, São Paulo, v.59, n.1, p.113-125, 2015.

- *RESUMEN: Se presenta la descripción de un método de extracción automática de candidatos a términos del área médica a partir del procesamiento de información lingüística. Para ello, se trabajó con reglas en el nivel léxico, morfológico y sintáctico. En primer lugar, se realizó la detección aplicando un diccionario estándar, el cual asignó a las palabras consideradas términos, la etiqueta MED (MÉDICO). Luego, para las palabras que no estaban contempladas en el diccionario (PNCD), se dedujeron las categorías gramaticales apelando a reglas morfológicas y sintácticas. Posteriormente, se procedió a la conformación de sintagmas nominales que involucraban PNCD y MED, para extraerlos como candidatos a términos del dominio. Se utilizaron los softwares Smorph y Módulo Post Smorph (MPS), que trabajan en bloque, y Xfst. Smoprh realiza el análisis morfológico y MPS trabaja sobre gramáticas locales. Xfst, por su parte, es una herramienta de estados finitos que opera sobre cadenas de caracteres, a las que asigna categorías previamente declaradas. El método se probó en una parte del corpus de casos clínicos compilado por Burdiles (2012), que contenía 217258 palabras, y los resultados arrojaron una precisión de 92,58%, una cobertura de 95,02% y una medida f de 93,78%.*

- *PALABRAS CLAVE: Terminología médica. Extracción automática. Información lingüística. Candidatos a término.*

REFERENCES

ABACCI, F. **Développement du module post-smorph.** 1999. Tesis (Maestría en Informática) – Memoria del DEA de Lingüística e Informática, Universidad Blaise-Pascal, Clermont-Ferrand, 1999.

AÏT MOKTHAR, S. **SMORPH:** guide d'utilisation: rapport technique. Clermont: Universidad Blaise Pascal: GRILL, 1998.

AMERICAN MEDICAL ASSOCIATION [AMA]. **United States adopted names council.** Disponible em: <http://www.ama-assn.org/ama/pub/physician-resources/medical-science/united-states-adopted-names-council.page>. Acceso en: 15 nov. 2013.

BEESLEY, K.; KARTTUNEN, L. **Finite state morphology.** Stanford: CSLI Stanford University, 2003.

BURDILES, G. **Descripción de la organización retórica del género caso clínico de la medicina a partir del corpus CCCM-2009.** 2012. 199p. Tesis Doctoral – Instituto de Literatura y Ciencias del Lenguaje, Facultad de Filosofía y Educación, Pontificia Universidad Católica de Valparaíso, Valparaíso, 2012.

CABRÉ, M. Morfología y terminología. In: FELÍU, E. **La morfología a debate.** Jaén: Universidad de Jaén, 2006. p.131-144.

CASTRO, E. et al. Automatic identification of biomedical concepts in Spanish language unstructured clinical texts. In: CASTRO, E. et al. In: ACM INTERNATIONAL HEALTH INFORMATICS SYMPOSIUM, 1., 2010, Nueva York. **Proceedings…** Nueva York: ACM, 2010. p.751-757.

DICCIONARIO esencial de la lengua española. Madrid: RAE, 2006.

DICCIONARIOS de términos médicos. Buenos Aires: Editorial Médica Panamericana, 2012.

DURUSSEL, B. **Terminología médica.** Santa Fe: Universidad Nacional del Litoral, 2006.

GÁLVEZ, C. Reconocimiento y anotación de nombres de fármacos genéricos en la literatura biomédica. **Acimed**, La Habana, v.23, n.4, p.326-345, 2012.

GARLA, V.; BRANDT, C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. **BMC Bioinformatic 2012,** Londres, v.13, n.261, 2012. Disponible en: <http://www.biomedcentral.com/1471-2105/13/261>. Acceso en: 30 nov. 2013.

IHTSDO. **SNOMED:** The global language of healthcare. Disponible en: <http://www.ihtsdo.org/snomed-ct/>. Acceso en: 15 nov. 2013.

JACQUEMIN, C.; BORIGAULT, D. Term extraction and automatic indexing. In: MITKOV, R. (Ed.). **The Oxford Handobook of Computational Linguistics.** Oxford: Oxford University Press, 2005. p.599-615.

KRAUTHAMMER, M.; NENADIĆ, G. Term identification in the biomedical literature. **Journal of Biomedical Informatics,** San Diego, v.37, n.6, 512-526, 2004.

LÓPEZ, C.; TERCEDOR, M., FABER, P. Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. **Revista E-Salud,** Málaga, v.2, n.8, p.228-240, 2006.

LÓPEZ-HUERTAS, M.; BARITÉ, M.; TORRES, I. Terminological representation of specialized areas in conceptual structures: the case of gender studie. In: LÓPEZ-HUERTAS, M.; BARITÉ, M.; TORRES, I. INTERNATIONAL ISCO CONFERENCE, 8., 2004, London. **Proceedings…** London: Ia C. McIlwaine, 2004. p.263-268.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval.** Cambridge: Cambridge University Press, 2009.

MARINCOVICH, J. Palabra y término: ¿Diferenciación o complementación?. **Revista Signos:** Estudios de Lingüística, Valparaíso, v.41, n.67, p.119-126, 2008.

MORENO-SANDOVAL, A. Panorama actual de la ingeniería lingüística. In: AMPARO, A.; RAMBLA, E.; VALERO, E. (Ed.). **Terminología y sociedad del conocimiento.** Berlín: Peter Lang Bern, 2009. p.99-116.

MORENO-SANDOVAL, A.; CAMPILLOS-LLANOS, L. Desing an annotation of multimedica: a multilingual text corpus of the biomedical domain. **Procedia:** Social and Behavioral Sciences, Amsterdam, v.95, p.33-39, 2013.

NATIONAL LIBRARY OF MEDICINE [NLM]. **Unified Medical Language System (UMLS).** Disponible en: <http://www.nlm.nih.gov/research/umls/>. Acceso en: 15 nov. 2013.

NOOJ. Disponible en: <http://www.nooj4nlp.net/pages/nooj.html>. Acceso em: 15 nov. 2013.

NUEVA gramática de la lengua española. Madrid: RAE, 2010.

SAGER, J. Pour une approche fonctionnelle de la terminologie. In: BÉJOINT, H.; THOIRON, P. (Ed.). **Le sens en terminologie.** Lyon: Presses Universitaires de Lyon, 2000. p.40-60.

SÁNCHEZ, D.; BATET, M.; VALLS, A. Web-based semantic similarity: an evaluation in the biomedical domain. **Int. J. Software and Informatics,** Beijing, v.4, n.1, p.39-52, 2010.

SEGURA, I.; MARTÍNEZ, P.; SAMY, D. Detección de fármacos genéricos en textos biomédicos. **Procesamiento del lenguaje natural,** Jaén, v.40, p.27-34, 2008.

TUASON, O. et al. Biological nomenclature: a source of lexical knowledge and ambiguity. In: PACIFIC SYMPOSIUM OF BIOCOMPUTING, 9., 2004, Oak Ridge. **Proceedings…** Oak Ridge: PSB, 2004. p.238-249.

VARELA, S. **Morfología lexica:** la formación de palabras. Madrid: Gredos, 2005.

VILLAYANDRE, M. **Aproximación a la lingüística computacional.** León: Universidad de León, 2010.

VIVALDI, J.; RODRÍGUEZ, H. Using Wikipedia for term extraction in the biomedical domain: first experiences. **Procesamiento del Lenguaje Natural,** Jaén, v.45, p.251-254, 2010.

Received december 2013

Accepted february 2014