Design of a Extraction System for Definitional Contexts from Biomedical Corpora

César Aguilar[α] and Olga Acosta[β]
[α]Pontificia Universidad Católica de Chile, Santiago de Chile
caguilara@uc.cl
[β]Cognitiva Latinoamérica, Santiago de Chile
oacosta@cognitiva.la

Abstract

In this paper we show a general advance about the desgin of a methodology for extracting definitional contexts from corpus of biomedicine in Spanish, taking into account a set of processes performed by the following modules: (i) a term extractor based in a hybrid method, (ii) a set of verbs that configure the syntactic structure of a definitional context, (iii) a chunker able to recognize those noun phrases that introduce a definition, considering the lexical relation of hyponymy/hypernymy, where the hyponym is the term defined, and the hypernym is the Genus Term which represents a conceptual category associated with such term.

1 Introduction

It is not surprising that, given the overwhelming amount of biomedical knowledge recorded in physical and electronic texts, currently there is an interest for developing semantics resources and tools oriented to improve the search and classification of biomedical concepts. Projects such as *Gene Ontology* [Smith *et al*., 2005], or *BioText Search Engine* [Hearst *et al*., 2007] are good examples of systems capable to extract and organize concepts, taking into account lexical-semantic relationships expressed in natural language.

Most of these projects have been developed for English, having in mind the big amount of documents produced. A paradigmatic example is **PubMed**, a search engine with accessing primarily the MEDLINE database of references and abstracts on biomedical topics. PubMed has been used in experiments oriented to the automatic classification of concepts extracted from large-corpora [Smith *et al*., 2005].

However, in Latin America, including Chile, there are no such projects in NLP. In order to fill this gap, we sketch here a method for extracting *definitional contexts* (abbreviated DCs), which are discursive structures that contain relevant information to define a term. A DC has at least three constituents: a term, a definition, and a verbal phrase that links both previous. Concurrently, we can identify other linguistic or metalinguistic units, whose function is to highlight the presence of a DC in a text, e.g.: discursive and typographical patterns [Sierra *et al.*, 2008; Acosta, Sierra and Aguilar, 2011]. An example is:

> [In general $^{\text{Discursive Pattern}}$], the [**paraprofessional workers** $^{\text{Term + Typographical Pattern}}$] [are defined as $^{\text{Verbal Phrase}}$] [those persons who are engaged in the provision of social care or social services, but who do not have professional training or qualifications $^{\text{Definition}}$]

According to this example, the term *paraprofessional workers* is emphasized by the use of bold font; the verbal phrase *are defined as* links the term *paraprofessional workers* to the actual definition *those persons who are engaged...* The term, the verbal phrase and the definition are discursive units introduced by the pragmatic pattern *in general.*

We conceive our method considering three central tasks:

- A term extraction that recognizes candidates to terms using a hybrid method based grammatical rules and stochastic techniques [Acosta, Aguilar and Infante, 2015].
- The use of a set of verbs that configure some specific kind of verbal phrase, called *predicative phrases* [Rothstein, 1983; Bowers, 1993; 2001], whose function is to link terms and definitions in a DC.
- The identification of lexical relations, particularly hyponymy/hyperonymy relations, in order to detect candidate to analytical (or Aristotelian) definitions, following the method proposed by Hearts [1992], Wilks, Slator and Guthrie [1996], as well as Acosta, Sierra and Aguilar [2011; 2015].
- Our paper is organized as follow: in the section 2 we describe in more detail the extraction of DCs from specialized corpora, attending the role of the predicative phrases (henceforth, PrPs) as grammatical linker among terms and definitions. Then, in section 3, we briefly explain our term extractor, and show some results generated searching biomedical terms in Spanish. In section 4, we show and describe a set of verbs that syntactically work as head of PrPs, and introduce analytical definitions in a DC. In section 5 we expose of methodology employed for identify

hyponyms and hyperonyms expressed in a biomedical Spanish documents, specifically situated in DCs.

2 Extraction of DCs

The development of methods and electronic tools for extracting conceptual information from texts has become an important task in NLP, mainly related with computational lexicography [Wilks, Slator and Guthrie, 1996], terminology [Malaisé, Zweigenbaum and Bachimont, 2005] and, in recent years, the building of ontologies [Navigli and Velardi, 2004; Velardi, Faralli and Navigli, 2013]. Reviewing in detail the criteria used to perform this type of extraction, we can recognize three ideas in common:

- Concepts are represented, in a natural language, by words, phrases or sentences. Thus, a definition is a linguistic structure useful for expressing this conceptual information [Sierra *et al*, 2008].
- If definitions are linguistic representations of concepts, then it is possible to recognize regular patterns in lexical, syntactic, semantic and discursive levels [Wilks, Slator and Guthrie, 1996].
- The use of statistical methods and computational tools for searching and extracting these regular patterns in large corpora. Therefore, the results are evaluated in order to determine if such patterns represent good or bad candidates to definitions [Malaisé, Zweigenbaum and Bachimont, 2005].

In line with these works and ideas, Sierra *et al*. [2008] delineate a method for recognizing and extracting terms and definitions expressed in DCs. As we have mentioned before, terms, PrPs and definitions configure the core of a DC, because these units show a recurrent use in specialized documents. Additionally, discursive and typographical patterns could be seen as optional units whose function is to introduce or indicate a potential DC in a text. We can represent the relation between all these units in this scheme:

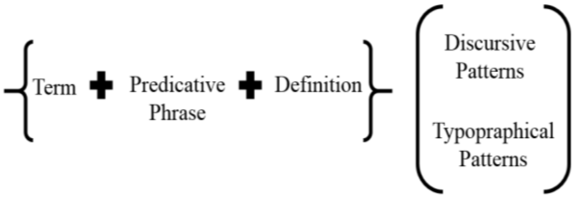


Figure 1, constitutive units of a DC structure

Having in mind this scheme, our proposal for extracting DCs in biomedical texts considers the identification of the main units, that is: terms, PrPs and definitions. Each unit in analyzed for a particular module, and the integration of all modules configures the architecture of our extraction system.

3 Term Extraction

We have developed a methodology for extracting single-word and multi-word terms from text-corpora, reported in Acosta, Aguilar and Infante (2015). Such methodology is supported for a hybrid approach, which including both a linguistic and a statistical phases.

In the linguistic part, the most frequent syntactic patterns are used to filter out candidate terms while, at the same time, removing non-relevant words from these candidates. In the statistical part, a corpus comparison approach is used to rank domain words [Kit and Liu, 2008]. A word occurring in both the reference and the domain corpus is ranked using relative frequency ratio [Manning and Schütze, 1999]. Given that words closely related with a domain should have a higher occurrence probability in that domain than in a reference corpus, we view a large reference corpus as an effective method for assigning relevance to domain words occurring in both corpora. If this ranking process is effective, the domain words will have higher weights than words not related to the domain.

For determining what word is a good candidate of term, we consider the notions of *termhood* and *unithood* proposed by Kageura and Umino [1996]. The *termhood* is described as the degree that a linguistic unit is related to domain-specific concepts. In contrast, the *unithood* refers to the strength of syntagmatic combinations and collocations which can be recognized as potential candidates to terms.

Thus, in the final stage, the word ranking can be used to extract multi-word candidate terms, so that words with high weights will contribute to increase the ranking of noun phrases when they are present (multi-word termhood). In the case of the *unithood*, we consider this to be assured in part for a syntactic filter [Vivaldi and Rodríguez, 2007] and the occurrence frequency of the noun phrase as a whole. Additionally, we propose implementing linguistic heuristics for automatically build a stopword list of non-relevant adjectives from the domain corpus. This latter is relevant since adjectives (primarily relational adjectives) have a compositional interpretation so that traditional measures (e.g., mutual information) fail in the task of showing the *unithood* of multi-word candidates.

We put attention in the terms represented for noun phrases (NPs) whose modifier is a relational adjective, because they assign a set of properties derived from an entity. In biomedical terminology, relational adjectives represent an important element for building specialized terms, e.g.: *inguinal hernia*, *venereal disease*, *psychological disorder* and others. For extracting these NPs with relational adjectives, we build a *chunker* that distinguishes the following patterns:

<RG><AQ>

<VAE><AQ>

<D.*|P.*|F.*|S.*><AQ><NC>

Where RG, AQ and VAE tags correspond to adverbs, adjectives and the verb *estar* (Eng. *To Be*), respectively. The tags <*D.**|*P.**|*F.**|*S.**> correspond to determinants, pronouns, punctuation signs and prepositions. The expression <*D.**|*P.**|*F.**|*S.**> is a restriction to reduce noise, since elements wrongly tagged as adjectives are extracted without this constraint. These tags are part of the system of annotation proposed for *FreeLing* (Carreras *et al*., 2004), which we have employed for tagging two corpora:

- A domain corpus composed for texts about human body diseases and related topics (surgeries, treatments, and so on) collected from MedlinePlus in Spanish. The size of this corpus is 1.2 million tokens.
- A reference corpus conformed for news and articles extracted from an online newspaper[1] from 2014. The size of this corpus is about 5 millions of tokens.

Using these chunker and patterns we perform an experiment for identifying terms, comparing whit four measures proposed by the following works:

- The *log-likelihood ratio* implemented by Gelbuk *et al*. [2010], abbreviated as LLR.
- The *word rank difference* employed by Kit and Liu [2008], abbreviated RD.
- The *relative frequence reason*, considered by Manning y Schütze [1999], abbreviated RFR.
- Finally, a *binomial approximation* using the *standard normal distribution* applied by Drouin [2003] for the *TermoStat* extraction system, abbreviated simply TS.

From a general point of view, in our experiment an important step is to eliminate the noise from terms removing the non-relevant adjectives automatically obtained from the domain corpus, as well as those words whose relative frequency in the reference corpus is greater than that in the domain corpus.

When we detect all the no-relevant adjectives, we generate a list as a filter for removing it, and then we can extract those NPs with relational adjectives.

Finally, once applied this filter, we obtained a precision of around 72.7%, considering the RFR measure, and the RD measure with 70.5%, specifically in the first 1000 candidates detected).

On the other hand, in the case of the global recall, we obtained proximally 73% also in the 1000 candidates. In the tables 1 and 2 we show the results of our experiment, contrasting precision and recall.

Table 1, percentages of precision in the extraction of terms using the adjective filter taken from reference corpus

	LLR	RD	RFR	TS
500	74.2	76.4	79	33.2
1000	66.4	70.5	72.7	28.9
1500	58.9	64.7	67.3	24.6
2000	53.9	64.5	60.7	18.7
2500	50.1	63.8	56.6	14.9
3000	48.4	60.1	53.8	12.4
3500	48.6	53.6	53.3	
4000	49.4	48.6	49.5	
4500	44.0	44.0	44.0	
5000	39.6	39.6	39.6	

Table 2, percentages of recall in the extraction of terms using the adjective filter taken from reference corpus

	LLR	RD	RFR	TS
500	16.5	17.0	17.5	7.4
1000	29.5	31.3	32.3	12.8
1500	39.2	43.1	44.8	16.4
2000	47.8	57.3	53.9	16.6
2500	55.6	70.8	62.8	16.6
3000	64.4	80.1	71.6	16.6
3500	75.5	83.3	82.9	
4000	87.6	86.3	87.8	
4500	87.8	87.8	87.8	
5000	87.8	87.8	87.8	

4 DCs and PrPs

In the case of PrPs, according to the analysis reported by Sierra *et al*. [2008], as well Aguilar, Acosta and Sierra [2010], these phrases configure the syntactic core of a DC. Syntactically, all PrP is structured around a relation *X-is-a-Subject-of/Y-is-a-predicate-of*. This relation is regulated by a syntactic rule named *rule of predicate linking*, proposed by Rothstein [1983]. This rule establishes a relation of saturation among the subject and the predicate, deriving two basic conditions:

I. X is the subject of the predicate of Y, if X is linked to Y.
II. If Y is the predicate of X, then Y cannot be predicated of anything else other than X.

Following Rothstein's explanation, Bowers [1993, 2001] develops a simple model to describe the syntactic configuration of these phrases. The PrP is mapped by a functional head, and its grammatical behaviour is similar to that of phrases such as Inflexional Phrase (IP) or Complement Phrase (CP).

Based on this description, we can infer two types of predicative phrases: a *primary predication*, i.e., those predicative phrases conformed by a subject to the left of the verb, and a predicate that is located to the right of the verb:

[Conjunctivitis [is [an inflammation of the conjunctiva of the eye $_{\text{NP}}$] $_{\text{PrP}}$] $_{\text{NP}}$]

[1] *La Jornada*. WEB site: www.lajornada.com.mx. Mexican newspaper with information available online.

In contrast, a *secondary predication* integrates a subject in a pre-verbal position, and an object and its predicate, both after the verb. In this case, the predicate affects the object of a sentence:

> [Watson and Crick [define [the DNA [as a molecule [that carries the genetic instructions used in the development, functioning and reproduction of all known living organisms $_{CP}$] $_{PrP}$]$_{NP}$]$_{VP}$]$_{IP}$]

A relevant difference observed in both examples is the explicit mention of the author(s) of the definition in the DC. According to Aguilar, Acosta and Sierra [2010], it is possible to determine two specific patterns:

(i) A pattern that follows the sequence Term + PrP + Definition, which is recognized as a **primary predication**.
(ii) Other pattern that follows the sequence Author + Term + PrP + Definition, which is recognized as a **secondary predication**.

Taking into account such kinds of PrPs, we can identify analytical definitions, assigning to its components, Genus Term and Differentia, a specific syntactic pattern. Thus, in the case of definitions associated to primary predications, the pattern is:

Table 3, construction pattern for primary predication linked to analytical definition

Definition	Genus Term	Differentia
Analytical (Primary PrP)	Noun Phrase = Noun + {AdjP/PP}*	CP = Relative Pronoun + IP
		PP = Preposition + NP
		AdjP = Adjective + NP

In contrast, in the case of analytical definitions related to secondary predications, the construction pattern is:

Table 4, construction pattern for secondary predication linked to analytical definition

Definition	Adverb/ Preposition	Genus Term	Differentia
Analytical (Secondary PrP)	*Como* *Por*	NP = Noun + {AdjP /PP}*	CP = Relative Pronoun + IP
			PP = Preposition + NP
			AdjP = Adjective + NP

The use of these patterns of PrPs for extracting terms and definitions has allowed to reach good results. For example: Sierra *et al.* [2008], as well as Alarcón, Sierra and Bach [2008] explored a specialized corpora about human genome and medicine (among others), integrated to the system BwanaNet —developed by the IULA-UPF[2]—, and they obtained a precision level around 0.58, and a recall of 0.83 for analytical definitions linked to verbs used in primary predications as *ser* (to be), *significar* (to mean/to signify), and also verbs used in secondary predications as *concebir* (to conceive) *definir* (to define), *entender* (to undestand), *identificar* (to identify), etc. Attending the individual score of these verbs, the most relevant are *concebir* (precision 0.71/recall 0.98) *definir* (precision 0.84/recall 0.98), contrasting whit others like *entender* (precision 0.36/recall 0.95), and *identificar* (precision 0.31/recall 0.90).

5 Hyponymy/hyperonymy extraction

The results of the extraction of DCs using PrPs allow to develop a method for recognize analytical definitions, focusing in the detection of the Genus Term introduces for the verb that works as a head of PrP. We face this task of detection taking into account the prototype theory proposed by Rosch and Lloyd [1978], applied to the description of categorization processes. Based on this theory, we can recognize a distinction among basic and subordinate categories: in the first case the single-word terms represented for nouns as *enfermedad* (disease), *corazón* (heart), *sistema* (system), etc., which represent basic categories, as opposed with the second case where multi-words terms represent subordinates categories: *enfermedad venérea* (venereal disease), *paro cardiaco* (heart atack), *sistema nervioso* (nervous system), and others.

We used this distintion (single-word *versus* multi-word) not only for identifying terms, but also hyponyms and hypernyms, attending the role of the relational adjectives and the preposition *de* (of/from). We formulate a set of possible term patterns recognizible in medical documents:

Table 5, Term patterns

Pattern	Example
Noun + Adjective (Spanish) Adjective + Noun (English)	Enfermedad cardiovascular Cardiovascular disease
Noun + Prepositional Phrase (Spanish)	Enfermedad de Alzheimer Alzheimer's disease
Noun + Noun	Diabetes mellitus
Acronyms	VIH HIV
Noun + Letter	Vitamina A Vitamin A
Letter + Noun	H Pylori

In our experiments for finding hyponyms and hypernyms, we only consider relational adjectives [Acosta, Aguilar and Sierra, 2013; Acosta, Sierra and Aguilar, 2011; 2015], exploring a corpus of medical texts in Spanish, with a size of 1.3 million of words, collected from *MedLinePlus*, the search engine of *PubMed*.

In order to identify patterns of NPs associated to hypernyms and hyponims, we develop an heuristic based on the detection of relational adjetives. Thus, we

[2] For more reference about BwanaNet, see the following link: http://bwananet.iula.upf.edu/index.htm

consider *H* as set of all single-word hyperonyms implicit in a corpus, and *F* the set of the most frequent hyperonyms in a set of candidate analytical definitions by establishing a specific frequency threshold *m*:

$$F = \{x \mid x \in H, freq(x) \geq m\}$$

On the other hand, *NP* is the set of noun phrases representing candidate categories:

$$NP = \{np \mid head(np) \in F, modifier(np) \in adjective\}$$

Subordinate categories *C* of a basic level b are those holding:

$$C^b = \{np \mid head(np) \in F, modifier(np) \in relational\text{-}adjective\}$$

Where modifier (*np*) representing an adjective modifier from a noun phrase *np* with head *b*. Returning with Rosch and Lloyd [1978], these subcategories show relevant differences respect to a basic level of categorization.

6 Desing a system for DC extraction

In the following section, we sketch our method for searching DCs, integrating in modules the tasks previusly exposed.

6.1 Methodology

We focus our efforts in analytical definitions, assuming that such definitions are the best source finding hyponymy-hyperonymy relations. Our method started to pre-processing a text corpus, in order to tokenize it. Then we annotate this corpus with POS tags, using the TreeTagger [Schmid, 1994].

Once made it, we employ syntactical and semantic filters for generating the first candidates of analytical definitions. The syntactical filter consists on a chunk grammar considering verb characteristics of analytical definitions, and its contextual patterns [Sierra *et al*., 2010], as well as syntactical structure of the most common constituents such as term, synonyms, and hypernyms.

On the other hand, the semantic phase filters candidates by means of a list of noun heads indicating relations part-whole and causal as well as empty heads semantically not related with term defined. An additional step extracts terms and hypernyms from candidate set.

In the case of the extraction of subordinate categories, we consider NPs with relational adjectives as modifiers of a term. The Figure 2 shows this process:

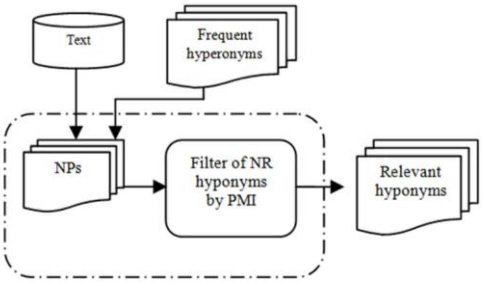


Figure 2, methodology for extracting subordinate categories

We obtain a set of NPs associated to relational adjectives and its frequency. Then, the NPs with hyperonyms as head are selected, and we calculate the pointwise mutual information (PMI) for each combination. Given its use in collocation extraction, we select a PMI measure, where PMI thresholds are established in order to filter non-relevant (NR) information. We considered the normalized PMI measure proposed by Bouma(2009):

$$i_n(x,y) = \left(\ln \frac{p(x,y)}{p(x)p(y)} \right) \Big/ -\ln p(x,y)$$

This normalized variant is due to two issues: to use association measures whose values have a fixed interpretation, and to reduce sensibility to low frequencies of data occurrence.

6.2 Corpus analysis and computational tools

As we have mentioned, our corpus is constituted for a set of medical documents, basically human body diseases and related topics (surgeries, treatments, and so on), collected from MedlinePlus in Spanish. Additionally, we use *NLTK* module [Bird, Klein and Loper, 2009], a set of open codes programming in Python language for analysing texts, in order to create a chunk parser for searching candidates to terms and hypernyms represented for NPs.

Integrating all the tasks exposed (the extraction of terms, the detection of PrPs associated to definitions, and the recognition of hyponyms/hypernyms), we conceive our methodology having in mind the following sequence of steps:

i) Processing a corpus and inserted POS tags for starting the extraction.
ii) Appliying the syntactic and semantics filters for generating candidates to DCs.
iii) We confirm the quality of these candidates if: (a) they are linked to a term linked to a PrP, and (b) they introduce a hyponymy/hyperonymy relation among the term and the Genus Term of a definition.

In the figure 3 we sketch our method:

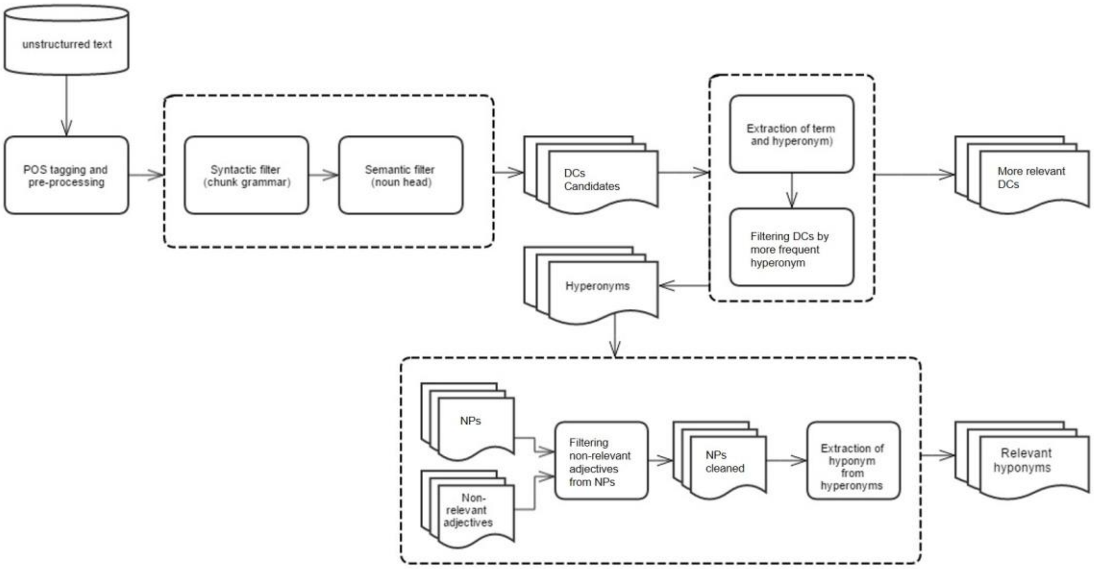

unstructurred text
POS tagging and pre-processing
Syntactic filter (chunk grammar)
Semantic filter (noun head)
DCs Candidates
Extraction of term and hyperonym)
Filtering DCs by more frequent hyperonym
More relevant DCs
Hyperonyms
NPs
Non-relevant adjectives
Filtering non-relevant adjectives from NPs
NPs cleaned
Extraction of hyponym from hyperonyms
Relevant hyponyms

Figure 3, architecture of prototype system for extracting DCs

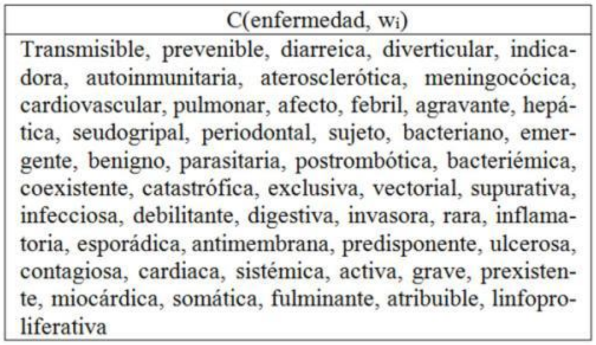
The architecture proposed here is an advance in the identification of DCs. According to the results reported by Acosta, Sierra and Aguilar [2015], the levels of precision and recall increase significantly when it is included the detection of hyponyms and hypernyms, in comparison to the results showed by Alarcón, Sierra and Bach [2008]:

Table 6, Comparison of results

	Precision	**Recall**
Alarcón, Sierra and Bach [2008]	41%	46%
Acosta, Sierra and Aguilar [2015]	62%	58%

Hypernyms, as generic classes of a domain, are expected to be related to a great deal of modifiers such as relational adjectives reflecting more specific categories (e.g., *cardiovascular disease*) than hyperonyms, or simply sensitive descriptions to a specific context (e.g., *rare disease*). In the table 7, we show the hypernym *enfermedad* (Ing. *disease*) and the first most related subset of 50 adjectives, taking into account its PMI values. In this example, only 30 out of 50 (60%) are relevant relations. In total, *disease* is related to 132 adjectives, of which 76 (58%) can be considered relevant:

Table 7, First 50 adjectives linked to the noun *enfermedad*



C(enfermedad, w_i)
Transmisible, prevenible, diarreica, diverticular, indicadora, autoinmunitaria, aterosclerótica, meningocócica, cardiovascular, pulmonar, afecto, febril, agravante, hepática, seudogripal, periodontal, sujeto, bacteriano, emergente, benigno, parasitaria, postrombótica, bacteriémica, coexistente, catastrófica, exclusiva, vectorial, supurativa, infecciosa, debilitante, digestiva, invasora, rara, inflamatoria, esporádica, antimembrana, predisponente, ulcerosa, contagiosa, cardiaca, sistémica, activa, grave, prexistente, miocárdica, somática, fulminante, atribuible, linfoproliferativa

7 Final considerations

In this paper we have delineate a method for extracting DCs from biomedical corpus in Spanish. Based on our preliminary results, we consider that we have achieved a considerable improvement taking into account the role of the hyponymy/hyperonymy relations as an important element to validate autentical analytical definitions expressed in DCs.

This consideration allows to observe a particular relation among syntactic structures and lexical-semantic information formulated in such definitions: on the one hand, it is not enough to search DCs based only syntactic sequences, although such structures can be considered as an interface for accessing such lexical-semantic information.

On the other hand, this task for recognizing hyponyms and hypernyms DCs ca be an important step for building ontologies based on text information, in line with the model proposed by Buitelaar, Cimiano and Magnini [2005]. The hyponymy/hyperonymy relation allows to infer a conceptual hierarchy between terms (in our case, situated in a biomedical domain), according to the categorization formulated by experts of a specific area. Although it is necessary to explore other lexical-semantic relations (e. g. synonymy of meronymy), we can start initially with the advances achieved by our methodology, in order to implement as well as possible our prototype system.

References

[Acosta, Sierra and Aguilar, 2011] Olga Acosta, Gerardo Sierra and César Aguilar. Extraction of Definitional Contexts using Lexical Relations. *International Journal of Computer Applications*, 34(6): 46-53, November 2011.

[Acosta, Aguilar and Infante, 2015] Olga Acosta, César Aguilar and Tomás Infante. Reconocimiento de términos en español mediante la aplicación de un enfoque de comparación entre corpus. *Linguamática*, 7(2):19–34, December 2015.

[Acosta, Aguilar and Sierra, 2015] Olga Acosta, César Aguilar and Gerardo Sierra. Extracting definitional contexts in Spanish through the identification of hyponymy-hyperonymy relations. In Jan Žižka and František Dařena (eds.), *Modern Computational Models of Semantic Discovery in Natural Language*, pages 48-70. IGI Global, Hershey, Pennsylvania, USA, 2015.

[Aguilar, Acosta and Sierra, 2010] César Aguilar, Olga Acosta and Gerardo Sierra. Recognition and extraction of definitional contexts in Spanish for sketching a lexical network. In Thamar Solorio and Ted Pedersen (eds.), *Proceedings of 1st young investigators workshop on computational approaches to languages of the Americas*, pages 109-116, ACL Publications, Stroudsburg, USA, 2010.

[Alarcón, Sierra and Bach, 2008] Rodrigo Alarcón, Gerardo Sierra and Carme Bach. ECODE: A Pattern Based Approach for Definitional Knowledge Extraction. In Elisenda Bernal and Janet DeCesaris (eds.), *Proceedings of the XIII EURALEX International Congress*, pages 923-928, IULA-UPF, Barcelona, España, 2008.

[Bird, Klein and Loper, 2009] Steven Bird, Ewan Klein and Edward Loper. *Natural Language Processing whit Python*. O'Reilly, Sebastropol, California, USA, 2009.

[Bowers, 2001] John Bowers. The syntax of predication, *Linguistic Inquiry*, 24(4):591-636, 1993.

[Bowers, 1993] John Bowers, Predication. In Mark Baltin and Chris Collins (eds.), *The Handbook of Contemporary Syntactic Theory*. Blackwell, Oxford, UK:299-333.

[Buitelaar, Cimiano and Magnini, 2005] Paul Buitelaar, Philipp Cimiano and Bernardo Magnini. *Ontology learning from text*. IOS Press, Amsterdam, The Netherlands, 2005.

[Drouin 2003] Patrick Drouin. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99-115, 2003.

[Gelbuk *et al*., 2010] Alexander Gelbukh, Grigori Sidorov, Eduardo Lavin, y Liliana Chanona. Automatic Term Extraction using log-likelihood based comparison with general reference corpus. In Christina Hopfe, Yacine Rezgui, Elisabeth Métais, Alun Preece and Haijiang Li (eds.), *Natural Language Processing and Information Systems. LNCS*, pages 248-255, Springer, Berlin, 2010.

[Hearst, 1992] Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 539-545, Nantes, France, ACL Publications, 1992.

[Hearst et al., 2007] Marti Hearst, Anna Divoli, Harendra Guturu, Alex Ksikes, Preslav Nakov, Michael Wooldridge and Jerry Ye. BioText search engine: beyond abstract search. *Bioinformatics*, 23(16): 2196-2197, August 2007.

[Kageura and Umino, 1996] Kio Kageura and Bin Umino. Methods of automatic term recognition: A review. *Terminology*, 3(2):259-289, . 1996.

[Kit and Liu, 2008] Chunyu Kit and Xiaoyue Liu. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2):204-229, 2008.

[Malaisé, Zweigenbaum, and Bachimont, 2005] Malaisé, Véronique, Zweigenbaum, Pierre and Bachimont, Bruno. Mining defining contexts to help structuring differential ontologies, *Terminology* 11(1):21-53, 2005.

[Manning and Schütze, 1999] Chris Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[Navigli and Velardi, 2004] Roberto Navigli and Paola Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30(2):151-179, 2004.

[Rosch and Lloyd, 1978] Eleanor Rosch and Barbara Lloyd. *Cognition and categorization*, Erlbaum, Hillsdale, New Jersey, 1978.

[Rothstein, 1983] Susan Rothstein, *The syntax forms of predication*, Ph. D. Thesis, MIT, Cambridge, Massachusetts, 1983.

[Schmid, 1994] Helmut Schmid. Probabilistic Part-of-Speech Tag-ging Using Decision Trees. In *Proceedings of International Conference of New Methods in Language*. Manchester, UK, 1994. WEB Site: www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

[Sierra et al., 2008] Gerardo Sierra, Rodrigo Alarcón, César Aguilar and Carme Bach. Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1):74–98, 2008.

[Smith et al., 2005] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6 (5):R-46, 2005.

[Velardi, Faralli and Navigli, 2013] Paola Velardi, Stefano Faralli and Roberto Navigli. OntoLearn Reloaded: A Graph-based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665-707, 2013.

[Vivaldi and Rodríguez, 2007] Vivaldi, Jorge, y Horacio Rodríguez. Evaluation of terms and term extraction systems: A practical approach". *Terminology*, 13(2):225-248, 2007.

[Wilks, Slator and Guthrie, 1995] Yorick Wilks, Brian M. Slator and Louise M. Guthrie. *Electric words*, MIT Press, Cambridge, Massachusetts, 1995.