Identification of Hyponyms, Hyperonyms, Meronyms and Holonyms in Medical Texts: a Cognitive Approach

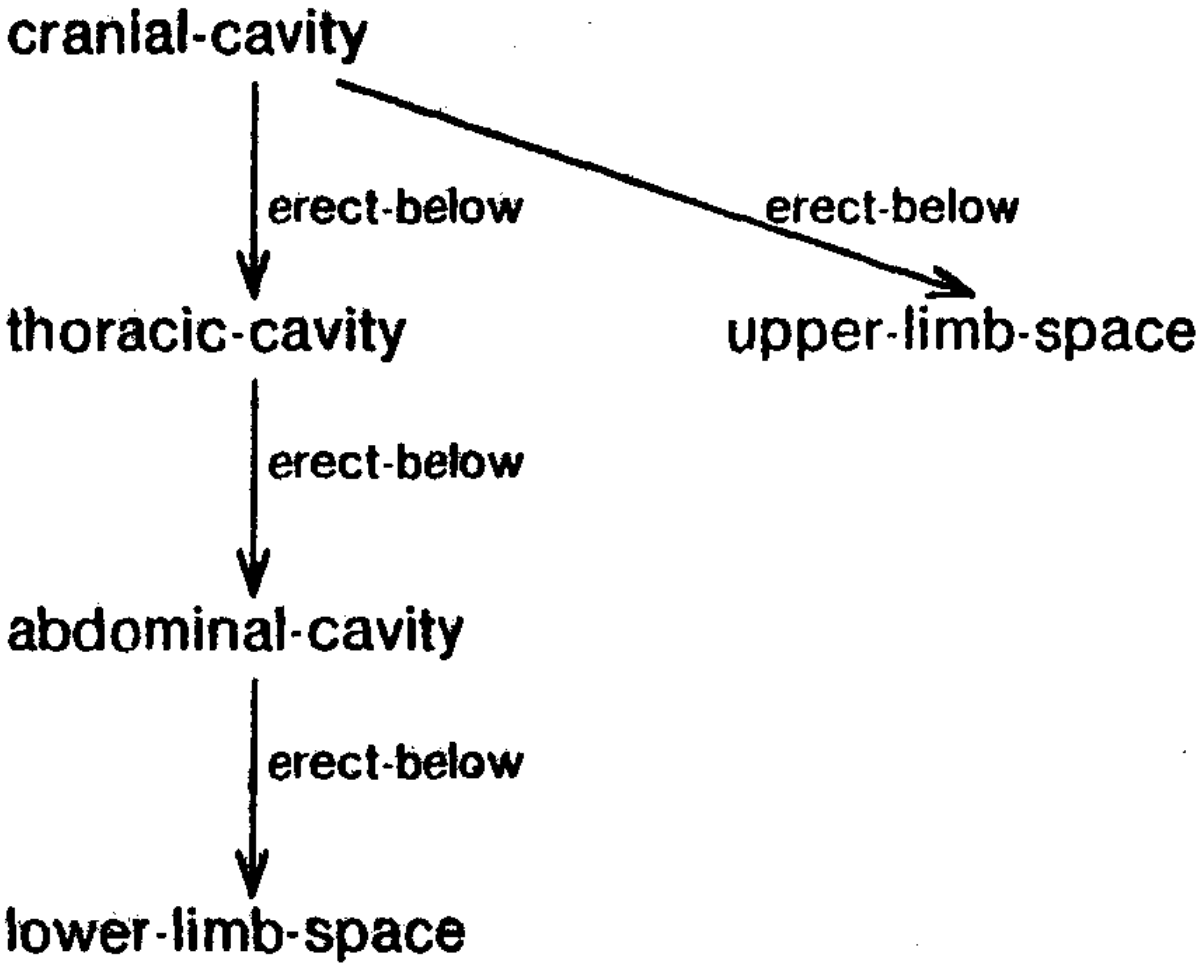
César Aguilar and Olga Acosta

caguilara@uc.cl
olgalimx@gmail.com

Cognitum 2015

Workshop on Cognitive Knowledge Acquisition and Applications
Buenos Aires, Argentina

Introduction (1)

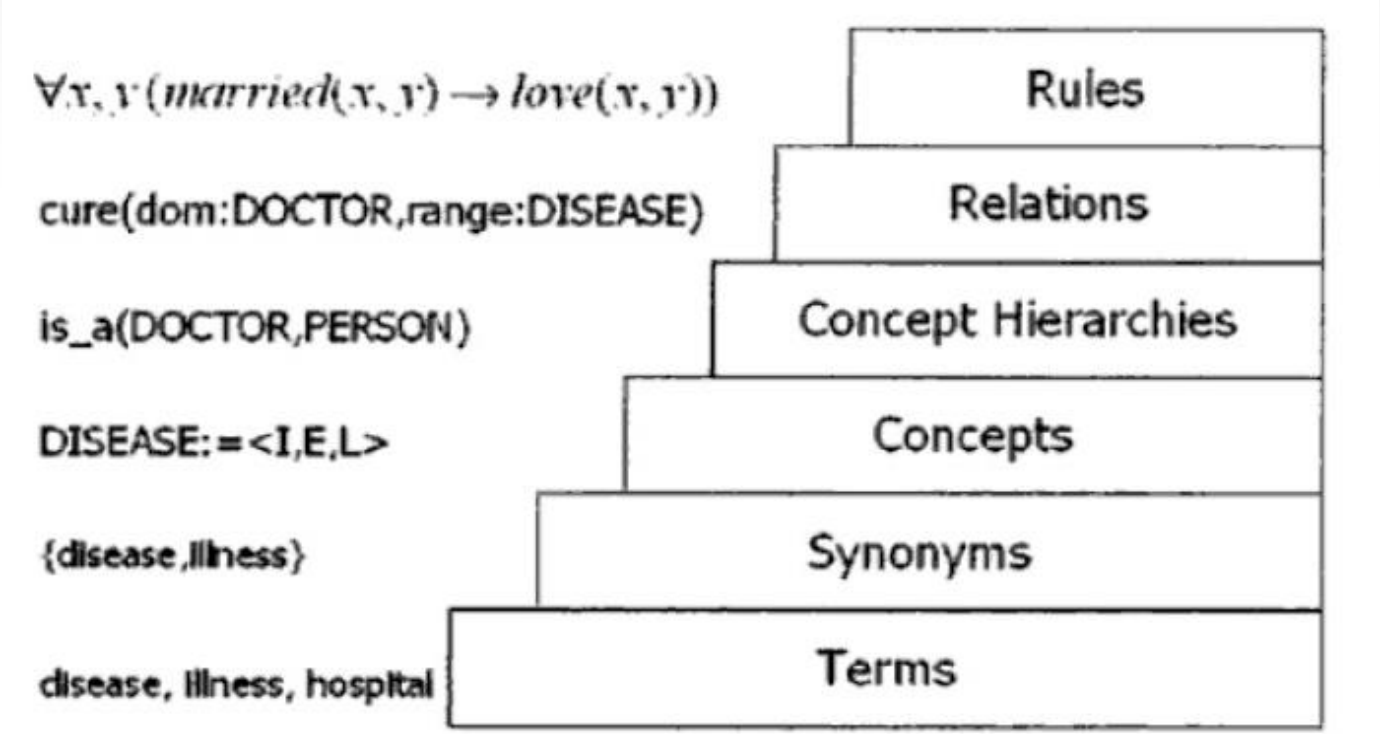


In this paper we propose two methods for extracting two types of lexical-semantic relations: (i) hyponymy/hyperonymy, (ii) meronymy/holonymy, from a medical corpus in Spanish.

In general, these methods consider the following steps: (a) the recognition of analytical definitional contexts in our corpus, (b) the identification of noun phrases (NP) that can represent hyponymy-hyperonymy or meronymy-holonymy relations taking into consideration only hyperonyms modified by relational adjectives.

Introduction (2)

The identification of lexical-semantic relations expressed in texts is the main goal of much of the research to date in NLP, particularly those oriented to the building of ontologies and taxonomies. A paradigmatic example of this kind of research is the volume prepared by Paul Buitelaar, Philipp Cimiano and Bernardo Magnini. These authors develop a complete methodology for building ontologies based on the extraction of conceptual information from text corpora.



Ontology learning layer cake according to Buitelaar, Cimiano and Magnani

Extraction of Hyponyms/Hyperonyms

Since the pioneering work of Hearst [1992], the most considered relation in this kind of extraction is the hyponym/hypernym. Based on Hearst's experiment, there are other alternative approaches:

Clustering: this approach emphasizes the distribution of context in corpus. According to this approach, words are characterized by its context and grouped by its similarity between contexts. One representative work about this approach is Faure and Nedellec [1998].

Finding patterns using the Web: in this approach new characteristic patterns and instances of the lexical relation of interest are extracted taking into account the Web as a huge source of textual information, according to the experiment performed by Pantel and Pennacchiotti [2006].

Machine learning: Finally, Snow, Jurafsky and Ng [2006] proposed an approach considering the application of machine learning methods, oriented to recognize useful patterns employing dependency paths.

Extraction of Meronyms/Holonyms

The first attempt for extracting automatically this kind of relation is the work of Matthew Berland and Eugene Charniak. They focused on genitive patterns identified through the employ of lexical seeds inserted in a part-whole relation with other words (e.g.: *the* ***basement*** *of a* ***building***). They use a news corpus of 100,000,000 words, and generate an ordered list of part-whole candidates inferred by a log-likelihood metric. Finally, they authors compare their results with the meronyms associated with **WordNet**, in order to determine precision.

On the other hand, Roxana Girju , Adriana Badulescu and Dan Moldovan conceived a different method considering a large list of possible patterns. They elaborated a corpora with texts taken from LA Times and Wall Street Journal (WSJ), and designed an algorithm named Iterative Semantic Specialization Learning (ISSL), which introduces a process of machine learning for recognizing new meronymy sequences in corpus. ISSL achieves a level of precision almost 83% in the LA Times, and 79% in the WSJ. In contrast, the level of recall is 79% for the first corpus, and 85% for the second corpus. Finally, they also compare the meronymy relations identified with a set of meronyms from WordNet.

Towards a cognitive perspective

Our methods of extraction take into account a cognitive perspective for recognizing lexical relations on texts. Particularly, we consider here:

Hyponymy/hypernymy relations

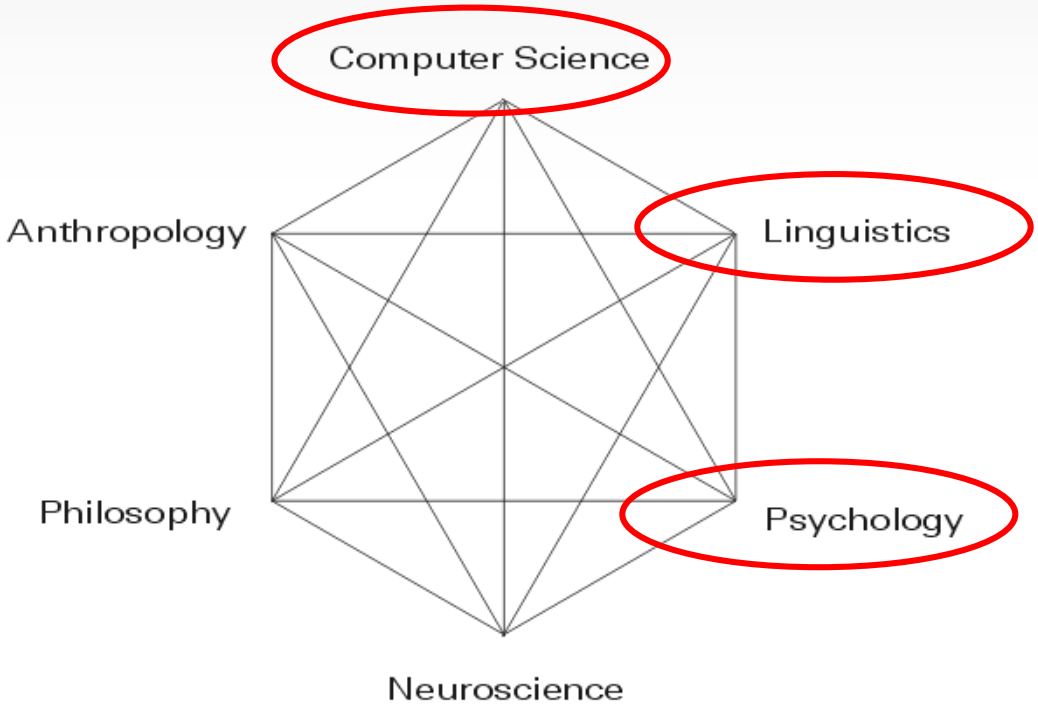
Categorization processes

Prototype theory

Meronymy/holonymy relations

Spatial scenes

Axial properties

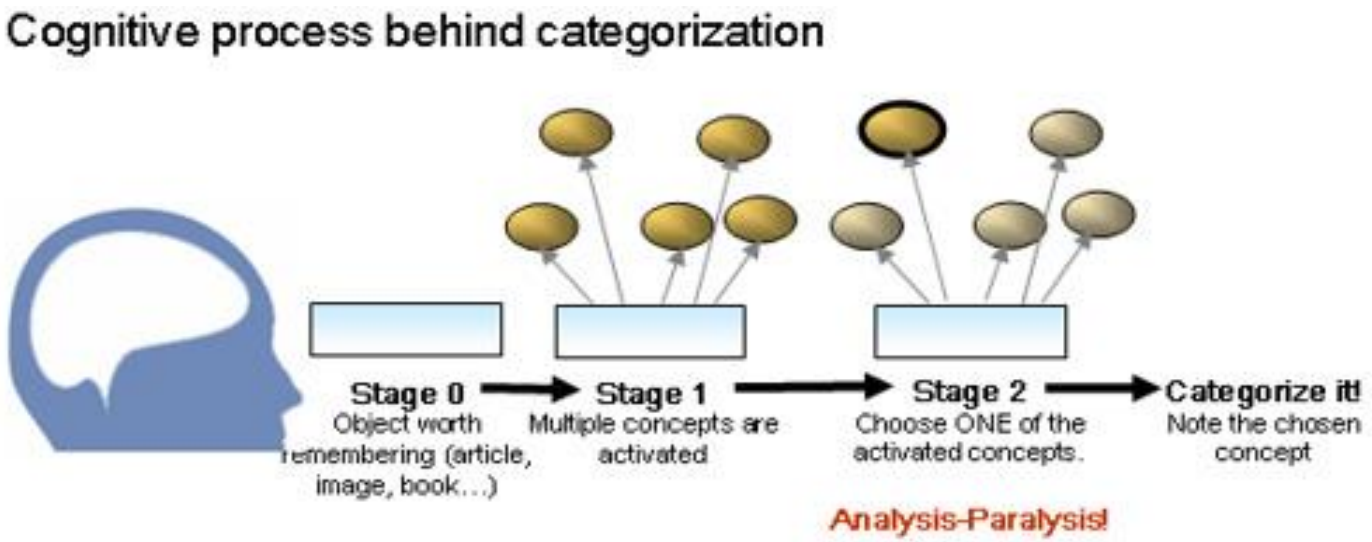


Categorization and concepts

Categorization is one of the most basic and important cognitive processes.

Categorization involves recognizing a new entity as part of abstract something conceived with other real instances

Concepts have a categorization function used for classifying new entities and extracting inferences about them.



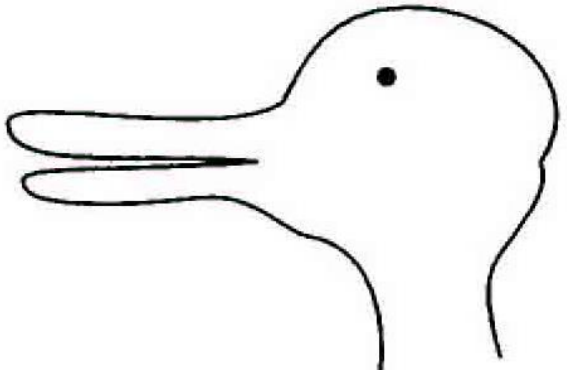
Principles of Categorization (1)

Rosch (1978) proposes two principles in order to build a system of categories.

The first refers to the function of this system, which must provide a maximum of information with the least cognitive effort.

The second emphasizes that perceived world (not-metaphysical) has structure. Maximum information with least cognitive effort is achieved if categories reflect the structure of the perceived world as better as possible.

Eleanor Rosch

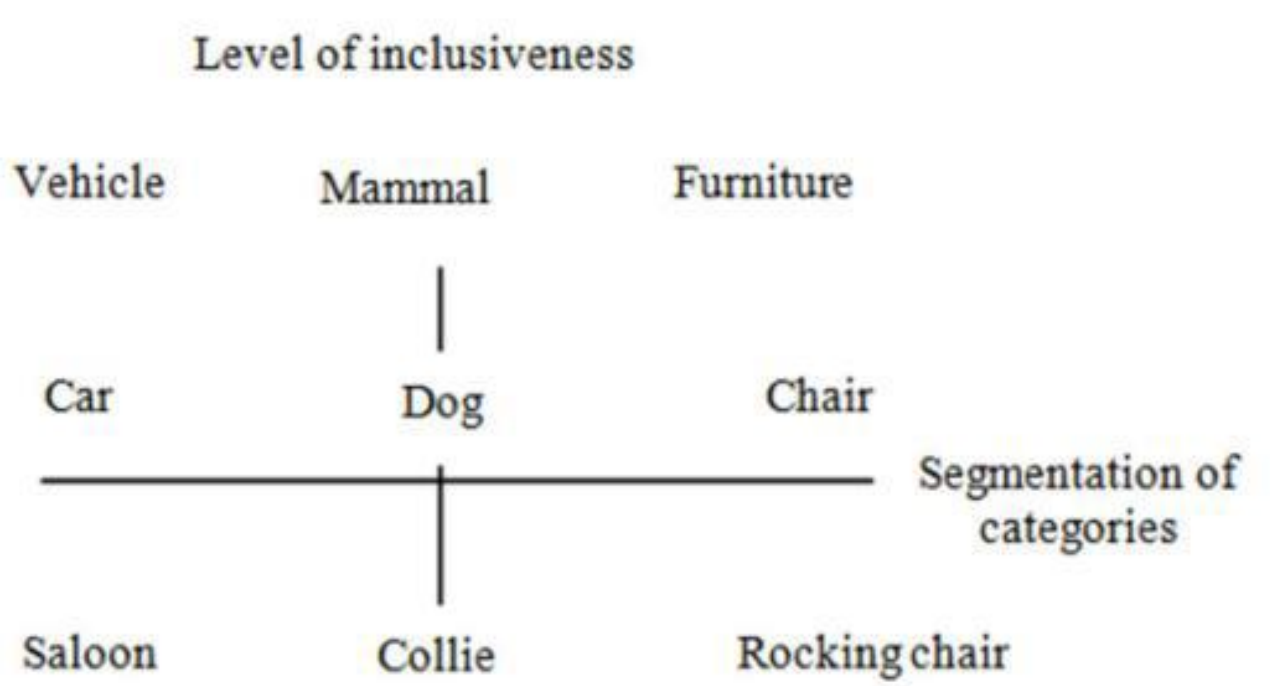


What it is?

Principles of Categorization (2)

In this figure, basic levels are associated with categories such as *car*, *dog* and *chair*. Categories situated on the top of the vertical axis (which provide less detail) are called *superordinate categories* (e.g.: *vehicle*, *mammal*, and *furniture*).

In contrast, those located in the lower vertical axis, which provide more detail, are called subordinate categories (e.g.: *saloon*, *collie*, and *rocking chair*).



Subordinate categories of our interest

Let *H* be 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*.

Defining types of adjectives

According to Violeta Demonte, in Spanish we can recognize two kinds of adjectives:

Descriptive adjectives: they refer to constitutive features of the modified noun. These features are exhibited or characterized by means of a single physical property: color, form, character, predisposition, sound, etc.: *el libro azu*l (the blue book), *la señora delgada* (the slim lady)

Relational adjectives: they assign a set of properties, e.g., all of the characteristics jointly defining names as: puerto marítimo (maritime port), paseo campestre (country walk). In terminological extraction, relational adjectives represent an important element for building specialized terms, e.g.: inguinal hernia, venereal disease, psychological disorder and others are considered terms in medicine.

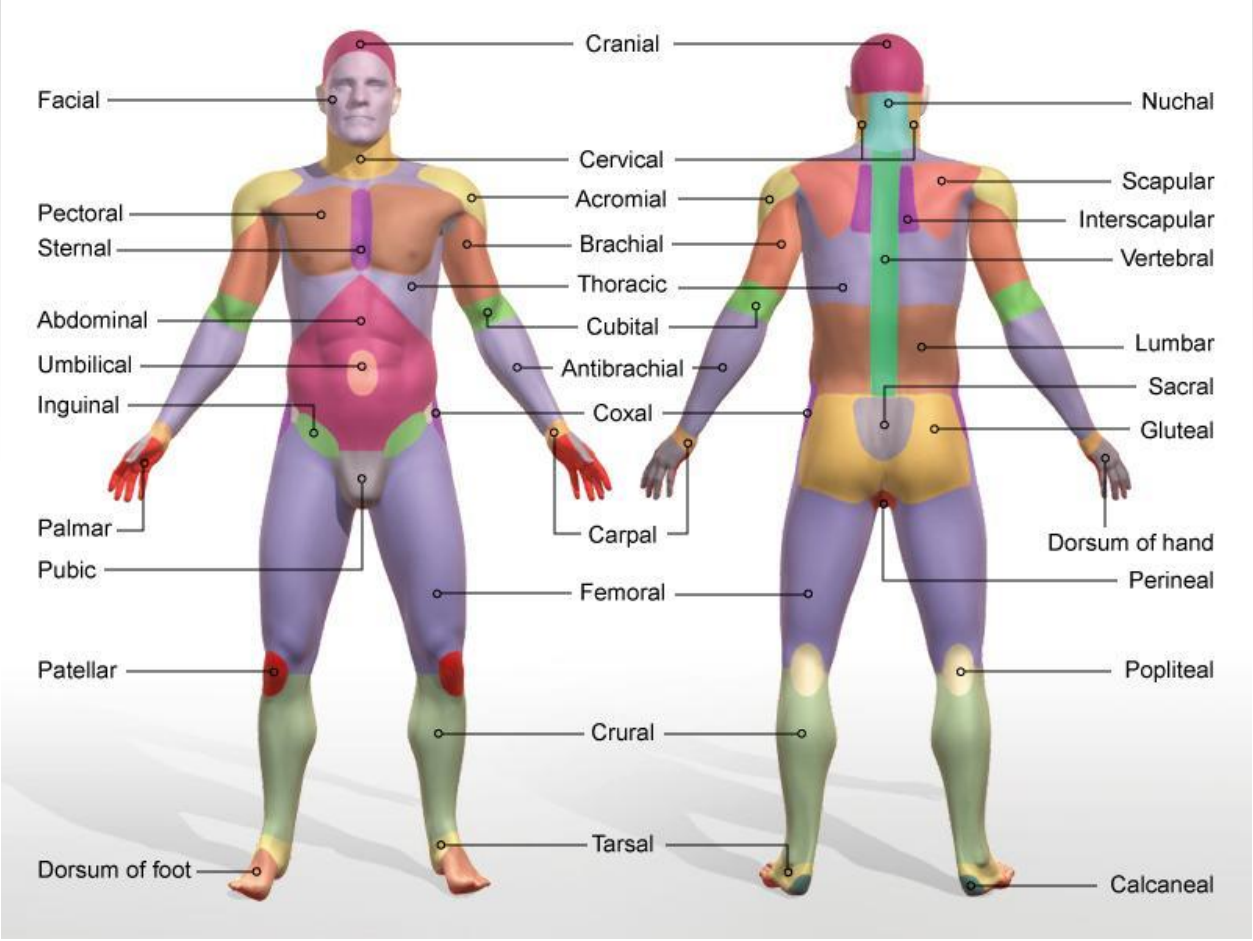
Spatial scenes

For our searching of meronyms and holonyms, we focus in the identification of concrete entities situated in spatial scenes. A spatial scene is a linguistic unit that contains information based on our spatial experience.

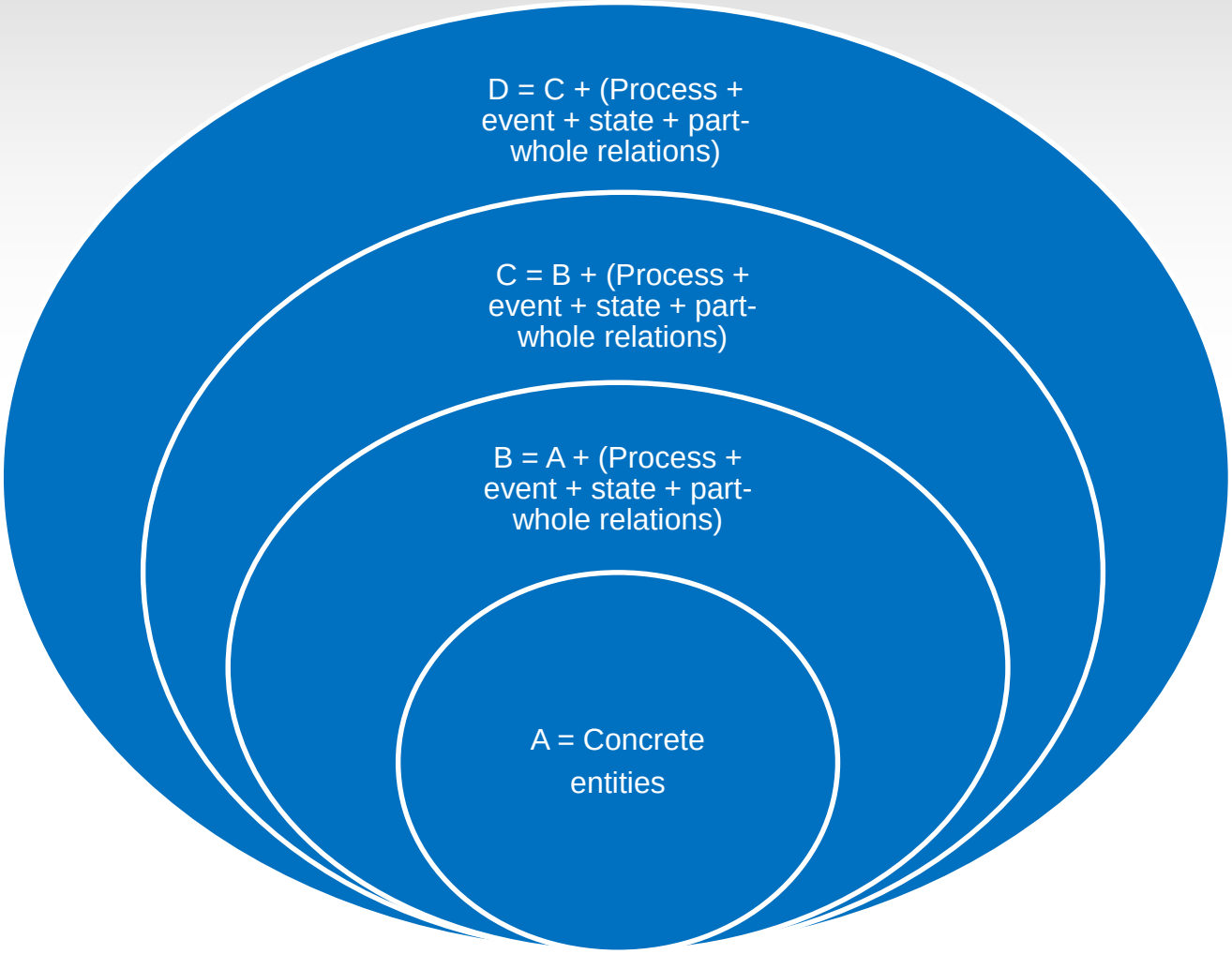
Examples of spatial scenes are: *the car is in front of the house*; *the cat is under the table*, *the bike is to the left of the tree*, and so on.

Concrete entities

Concrete entities potentially are involved in part-whole relations, as well as they can participate in processes, states, events, and so on. Linguistically, this behavior in Spanish can be represented by the preposition *de* (Engl., of/from): *finger of the hand*, *inflammation of the eyes*, *infection of the kidney*, *tumor of the breast (breast tumor)*, etc.



The productive use of preposición *de* (“of/from”)

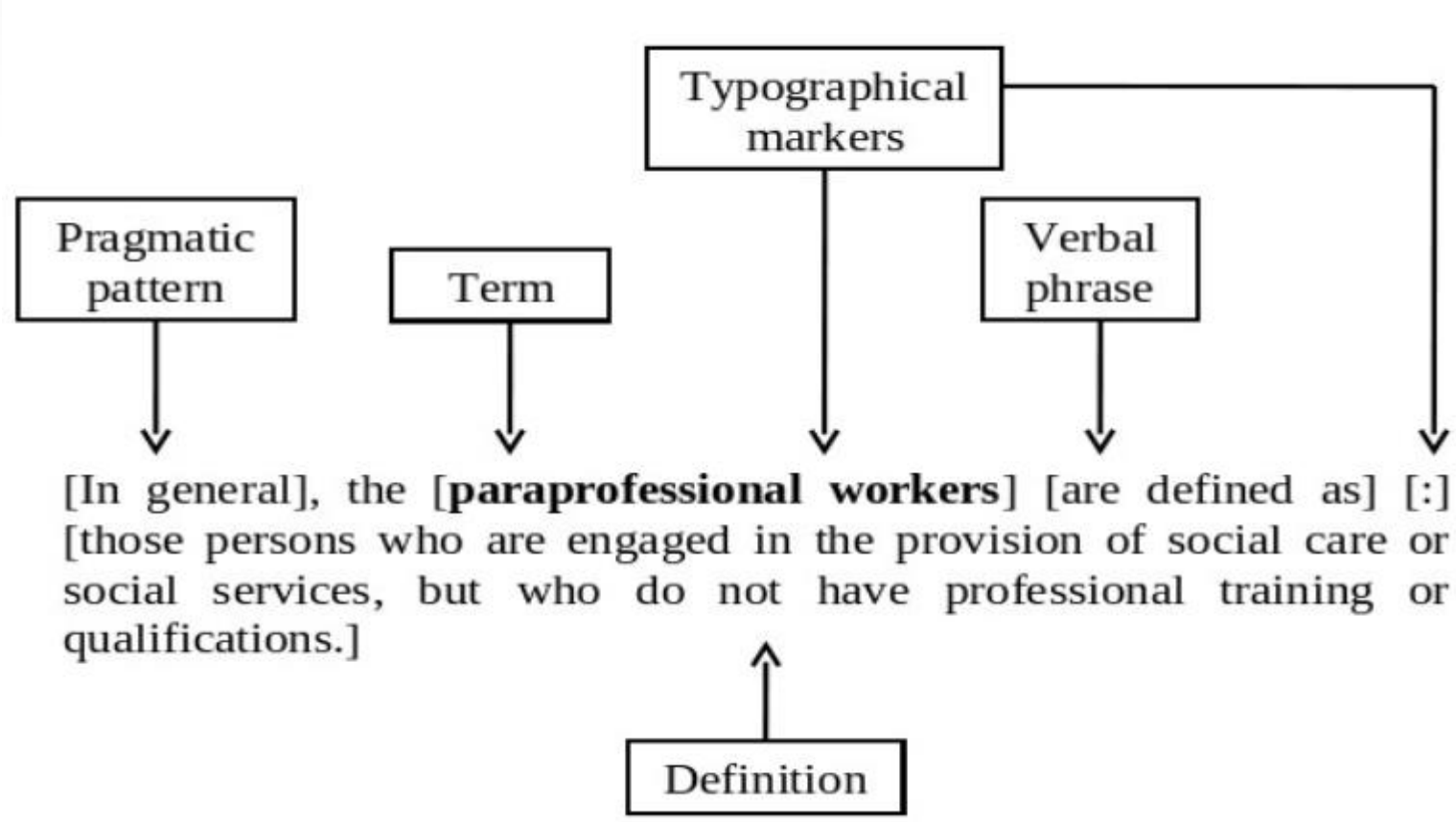


The most commonly used preposition in Spanish is *de*. Over 80% de noun phrases with prepositional phrase has as core *de*.

We hypothesized that *de* is more productive than other prepositions because represents several important kinds of semantic relations.

Searching definitional contexts (1)

For performing both search methods, firstly we focus to identify *definitional contexts* (or DCs) extraction. In brief, a DC is a discursive structure that contains relevant information to define a term. The DC has at least two constituents: a term and a definition, and usually linguistic or metalinguistic forms, such as verbal phrases, typographical markers and/or pragmatic patterns. An example is:



Searching definitional contexts (2)

We start this work from the traditional model of the definition formulated by Aristotle, the Analytical Definition composed of two units:

I. A *Genus term* is more general than the headword, and is related to it via an IS-A relation.

II. The *Differentia* is the word or the set of words that serves to differentiate the headword from other headwords with the same genus.

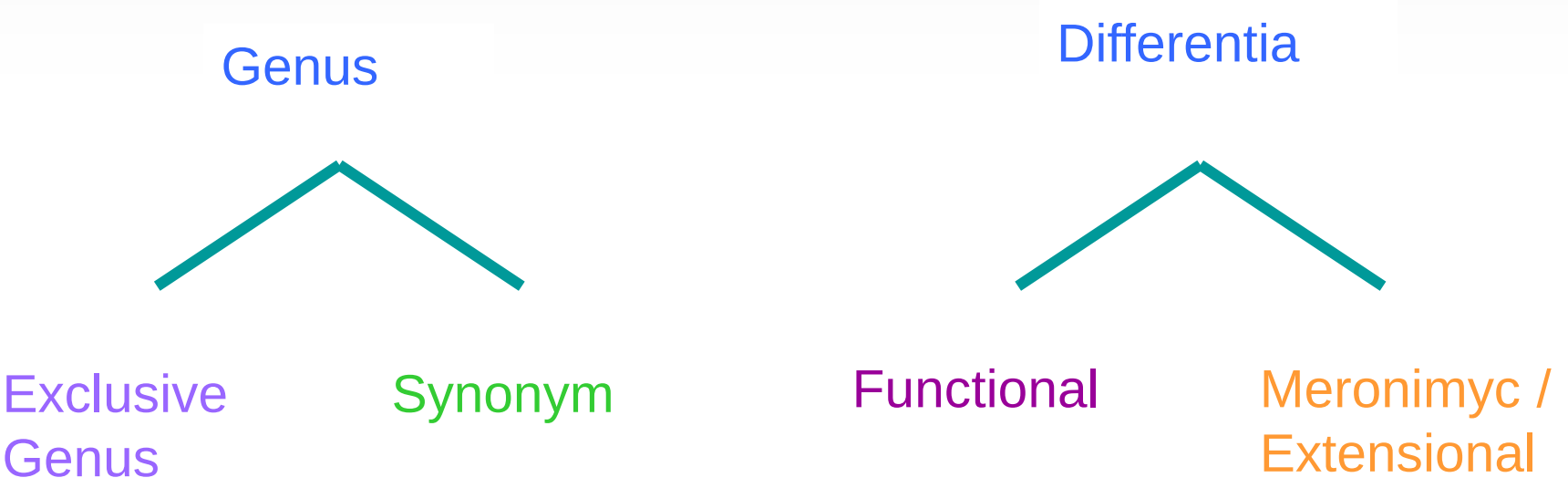
Genus term

Knife — a blade fixed in a handle, used for cutting as a tool or weapon.

Differentia

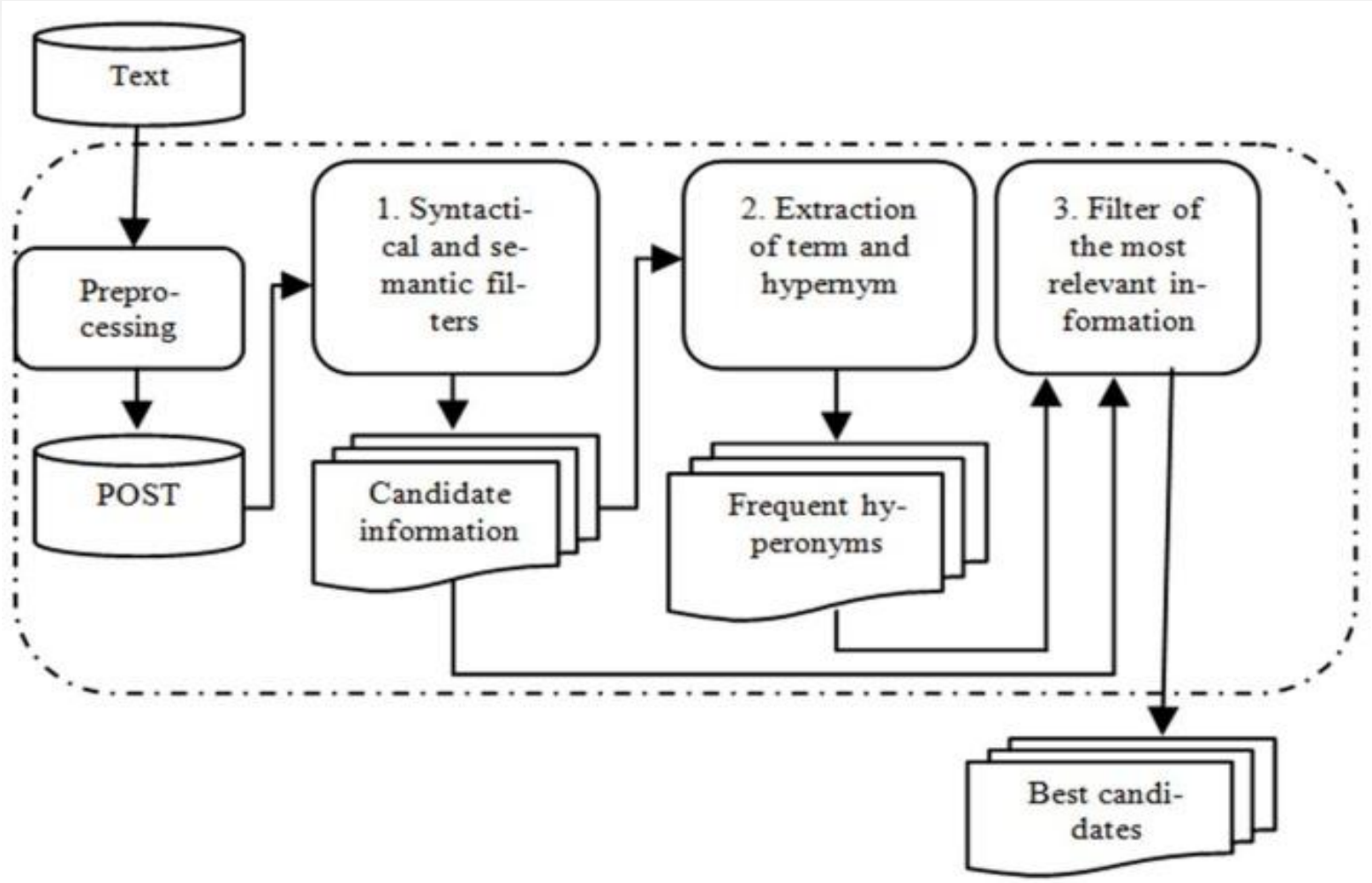
Searching definitional contexts (3)

Keeping in mind the Analytical Definition, we developed the following typology:



Extracting definitions from medical texts (1)

We assume that the best sources for finding hyponymy-hyperonymy relations are the definitions expressed in specialized texts. In order to achieve this goal, we develop a general methodology whose input are non-structured texts:



Extracting definitions from medical texts (2)

The previous figure shows the following processes:

1. The text source is tokenized in sentences, annotated with POS tags and normalized.

2. There are two syntactical and semantic filters which provide the first candidate set of analytical definitions.

3. Syntactical filter consists on a chunk grammar that analyzes the syntactic components of analytical definitions.

4. Semantic filters recognize 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.

5. An additional step extracts terms and hyperonyms from candidate set.

Extraction of subordinate categories

In the case of terms, we consider relational adjectives are used for building subordinate categories in specialized domains. We use the most frequent hyperonyms for extracting these relevant subordinate categories.

We obtain a set of noun phrases with structure: noun + adjective from corpus, as well as its frequency. Then, noun phrases with hyperonyms as head are selected, and we calculate the pointwise mutual information (PMI) for each combination. 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 fundamental issues: to use association measures whose values have a fixed interpretation, and to reduce sensibility to low frequencies of data occurrence.

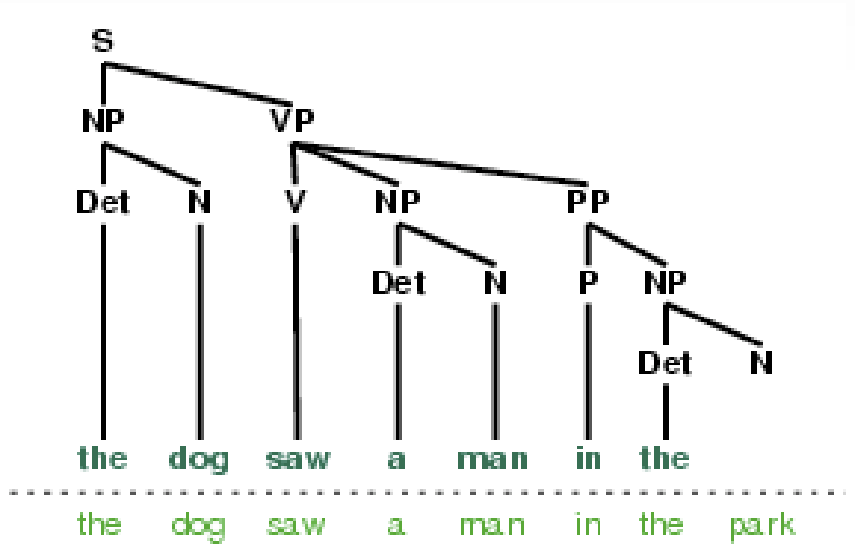
Analysis and results

Corpus: a set of documents collected from MedLinePlus in Spanish. Size of corpus: 1.2 million words. This corpus is tagged with POS.

Linguistic analysis tools: Natural Language Tool-Kit (NLTK)

Natural Language Toolkit

www.nltk.org



Results for hyponyms/hyperonyms (1)

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 this table we show the disease hypernym and the first most related subset of 50 adjectives, taking into account its PMI values. In this example extracted of a real corpus, 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.:

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, preexistente, miocárdica, somática, fulminante, atribuible, linfoproliferativa

Results for hyponyms/hyperonyms (2)

On the other hand, we can compare de degree of specialty between the a relational adjective as *cardiovascular*, and a descriptive adjective as *rare*:

C(w_i, cardiovascular)	C(w_i, raro)
efecto, problema, congreso, función, evento, relación, examen, inestabilidad, trastorno, enfermedad, bypass, causa, beneficio, sistema, reparador, descompensación, cirugía, operación, mortalidad, aparato, educación, síntoma, eficiencia, episodio, riesgo, investigación, manifestación, afección, medicamento, director, muerte, salud	televisión, enfermedad, complicación, infancia, niño, color, obesidad, mhc, nucleótido, sustancia, mutación, trastorno, grupo, meconio, epistaxis, derecha, síndrome, cáncer, alelo, forma, caso, párpado

Results for hyponyms/hyperonyms (3)

In order to face the phenomenon of compositionality between hyperonyms and relational adjectives that affect the performance of traditional measures, we automatically extract a stop-list of descriptive adjectives from the same source of input information, implementing three criteria proposed in Demonte (1999) for distinguishing between descriptive and relational adjectives. These criteria are:

- Adjective used predicatively: *the method is important*
- Adjective used in comparisons, so that its meaning is modified by adverbs of degree: *relatively fast*
- Precedence of adjective respect to the noun: *a serious disease*

Results for hyponyms/hyperonyms (4)

We consider two approaches based on patterns, and a baseline derived from only most common verbs used in analytical definitions. Both of the methods outperformed baseline's precision, but recall was significantly decreased.

On the one hand, the method proposed by Sierra et al. (2008) achieved a good recall (63%), but the precision was very low (24%). On the other hand, with the method proposed by Acosta et al. (2011) we achieved a high precision (68%), and a trade-off between precision and recall (56%).

	Precision	Recall	F-Measure
Baseline	8%	100%	15%
Sierra *et al.* (2008)	24%	63%	35%
Acosta *et al.* (2011)	68%	56%	61%

Results for hyponyms/hyperonyms (5)

We extract a set of descriptive adjectives by implementing linguistic heuristics. Our results show a high precision (68%) with a recall acceptable (45%). This subset of descriptive adjectives is removed from the set of noun phrases with structure: noun + adjective before final results. In this table we show the initial precision, that is, precision obtained without some filtering process:

Hyperonym	Initial precision	Hyperonym	Initial precision
Enfermedad	61	Tratamiento	36
Desorden	80	Cirugía	67
Examinación	52	Método	37
Condición	61	Problema	62
Procedimiento	40	Proceso	47
Infección	56	Inflamación	58
Proteína	67	Glándula	95
Cáncer	55	Órgano	43
Tumor	63	Medicamento	60

Results for hyponyms/hyperonyms (6)

These results show a significant improvement in precision from PMI 0.25, but recall is negatively affected as this threshold is increased.

On the other hand, if we consider linguistic heuristics we obtain a trade-off between precision and recall:

Hyperonym	Linguistic Heuristics		
	P	R	F
Enfermedad	79	74	76
Desorden	95	69	80
Examinación	74	85	79
Condición	88	75	81
Procedimiento	85	89	87
Infección	84	76	82
Proteína	88	76	82
Cáncer	69	94	80
Tumor	82	86	84
Tratamiento	68	70	69
Cirugía	90	82	86
Método	72	82	77
Problema	83	67	74
Proceso	73	73	73
Inflamación	82	78	80
Glándula	100	100	100
Órgano	71	71	71
Medicamento	76	72	74

Results for meronyms/holonyms (1)

For performing our search of meronyms and holonyms, we delineate a *chunking* process for detecting those NPs that expressed concrete entities linked to the preposition *de*. We formulate the following pattern:

<RG><PDEL><DA>?<NC><AQ>*

Where <DA> is a determinant tag and <PDEL> is a tag including contraction *del* and *of* plus an article (i.e., *de la*).

Before extracting concrete entities, the non-relevant set of automatically obtained nouns and adjectives are removed from the set of the phrases:

<NC>+<AQ>*<PDEL><NC><AQ>*
(<NC>+<AQ>*<VAE>)?<RG><PDEL><DA>?<NC><AQ>*

Results for meronyms/holonyms (2)

The first phase of extraction of concrete entities had a high precision of 73%. In a next *bootstrapping* step precision decreased to 51%. Next phases (3 to 5) hold measures without major changes in precision. In relation to the candidate Part-Whole relations were only extracted with the set obtained of the phase 1. Given the low precision obtained in phase 1 (24%), candidates of remaining phases were not analyzed. In general terms, recall achieved in the extraction of concrete entities was 65%.

We considered that one of the relevant results of our experiment, which was not addressed in the initial design, is the precision in recognition of terms of the domain. As it can be observed in the table 2, precision is over a 70%, which is a relevant result. On the other hand, we noted extraction of concrete entities tend to converge in a phase 5

Phase	Precision	Concrete entities	Candidates	Terms
1	73%	308	423	74%
2	51%	566	1109	74%
3	46%	744	1632	72%
4	44%	806	1849	71%
5	43%	821	1891	71%

Muchas gracias

Para mayor información:

CAguilar@iingen.unam.mx