Description and Evaluation of a Definition Extraction System for Spanish language

Rodrigo Alarcón
Universidad Nacional Autónoma de México, Grupo de Ingeniería Lingüística
Torre de Ingeniería, Basamento 3,
Mexico City
ralarconm@iingen.unam.mx

Gerardo Sierra
Universidad Nacional Autónoma de México, Grupo de Ingeniería Lingüística,
Torre de Ingeniería, Basamento 3,
Mexico City
gsierram@iingen.unam.mx

Carme Bach
Universitat Pompeu Fabra, Grupo IULATERM, Departament de Traducció i Ciències del Llenguatge,
Roc Boronat, 138, Barcelona, Spain
carme.bach@upf.edu

Abstract

In this paper we present a description and evaluation of a pattern-based approach for definition extraction in Spanish specialised texts. The system is based on the search for definitional verbal patterns related to four different kinds of definitions: analytical, extensional, functional and synonimical. This system could be helpful in the development of ontologies, databases of lexical knowledge, glossaries or specialised dictionaries.

Keywords

Definition extraction, definitional contexts, definitional verbal patterns, pattern-based approach.

1. Introduction

There is a growing interest in the development of systems for the automatic extraction of information that describe the meaning of terms. This information occurs in structures commonly called *definitional contexts* (DCs), which are structured by a series of lexical and metalinguistic patterns that can be automatically recognised. In this context, in this paper we present a work focused on developing a system for the automatic extraction of definitional contexts on Spanish language specialised texts. This system looks for instances of definitional verbal patterns, filters non-relevant contexts, identifies the main constituent elements on the candidates, i.e., terms and definitions, and performs an automatic ranking of the results.

Firstly, we will describe the structure of DCs; secondly, we provide a short review of related works; we then present the methodology followed for the automatic extraction of DCs together with an evaluation of this methodology; and lastly, we propose some future work.

2. Definitional Contexts in Specialised Texts

A definitional context is a textual fragment from a specialised text where a definition of a term is given. Its basic structure consists of a term (T) and its definition (D), both elements being connected by typographic or syntactic patterns. Typographic patterns are punctuation marks (comas, parenthesis), while syntactic patterns include definitional verbs –such as *definir* (to define) or *significar* (to signify)– as well as discursive markers –such as *es decir* (that is, lit. (it) is to say), or *o sea* (that is, lit. or be-subjunctive)–. Apart from these, DCs can include pragmatic patterns (PPR), which provide conditions for the use of the term or clarify its meaning, as in *en términos generales* (in general terms) or *en este sentido* (in this sense). For example:

"Desde un punto de vista práctico, los opioides se definen como compuestos de acción directa, cuyos efectos se ven antagonizados estereoespecíficamente por la naloxona."

In this case, the term *opioides* is connected to its definition (*compuestos de acción directa […]*) by the verbal pattern *se definen como* (are defined as), while the general sense of the context is modified by the pragmatic pattern *desde un punto de vista práctico* (from a practical point of view).

3. Advances in Definitional Contexts Extraction

Definition extraction from specialised texts has become a relevant task in the field of information extraction. In order to extract definitional information, the most common strategy is to extract certain recurrent patterns, which are commonly found in DCs.

The use of this kind of patterns has been applied on different scenarios. One of the first descriptive works can be found in [1], in which the behaviour of the contexts where terms occur is described. This work states that, when authors define a term, they usually employ typographic patterns to visually highlight the presence of terms and/or definitions, as well as lexical and metalinguistic patterns connecting DCs elements by means of syntactic structures. [2] reinforces this idea was reinforced and also explained the fact that definitional patterns can provide keys for the identification of the type of definition occurring in DCs, which facilitates the task of ontology development.

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 7–13,

Regarding applied works, [3] reports a system called *Definder* for the automatic extraction of definitions from medical texts in English. In the same line of research, other works have been focused on DCs extraction from specialised texts in other languages, for example German [4], Portuguese [5] or Spanish [6]. Definition extraction has also been used as a previous step for the automatic extraction of semantic relations or the automatic development of ontologies [7], [8], as well as for obtaining knowledge for the development of *eLearning* technologies [9].

Furthermore, the automatic extraction of definitions has been focused on direct Web exploitation. That is the case of the work reported in [10] whose main goal is the extraction of definitions from on on-line sources for question answering systems. [11] reports an application called *GlossExtractor*, that works on the Web, mainly online glossaries and Web specialised documents, also for the automatic extraction of definitions, but starting from a list of predefined terms. [12] developed a system called *DefExplorer* for definition extraction of Web documents for the Chinese Language.

All of these systems start from the search of specific definitional patterns in each language and they also integrate procedures for filtering non-relevant contexts, i.e., contexts that contain a definitional pattern that does not yield an actual definitional context. Finally, all of these methodologies are based on the exploitation of specialised documents, being the direct Web exploitation a recently incorporated process.

4. ECODE

As we have mentioned before, the main purpose of a definitional context extractor is to simplify the search of relevant information about terms, by means of searching for occurrences of definitional patterns.

An extractor that only retrieves those occurrences of definitional patterns would be a useful system for terminographical work. However, the manual analysis of the retrieved occurrences would still imply an effort that could be simplified by an extractor that includes the automatic processing of the obtained information.

Therefore, we propose a methodology that includes not only the extraction of occurrences of definitional patterns, but also a filtering process of non-relevant contexts (i.e. non definitional contexts), the automatic identification of the possible constitutive elements of a DC: terms and definitions, and a final automatic ranking of the results. This system is called *ECODE: extractor de contextos definitorios* (definitional contexts extractor).

A general overview of the system is shown in figure 1. It can be seen that the system input consists of a corpus tagged with POS categories, since some of them are necessary in the different processes of the system. It can also be seen that the main three processes are: a) the extraction of DC candidates, b) the analysis of DC candidates, and c) the evaluation of DC candidates.

The extraction of DC candidates is a process that uses a grammar of verbal patterns with some specific parameters: the definitional verbs to search for and the nexus that can also be part of the pattern, i.e., the adverb *como* (as) in the pattern *se define como* (it is defined as). In this case, the grammar shall also include constraints on the verbal times and grammatical person in which each verb can occur, as well as the different positions for each verb where the term can occur in a DC.

Once the DC candidates are extracted, they are analysed in the next process, which is carried out in two steps: the filtering of non-relevant candidates, and the identification of their constituent elements. The filtering process makes use of a set of linguistic and contextual rules to determine those cases where no DCs are found, while the identification of their constituent elements makes use of a decision tree, which also analyses the grammar of verbal patterns in order to identify the term and its definition on each DC candidate.

Finally, the system performs an automatic ranking of the candidates proposed as DCs. This process use a set of heuristic rules and aims to identify those candidates that follow a prototypical structure of terms and definitions.

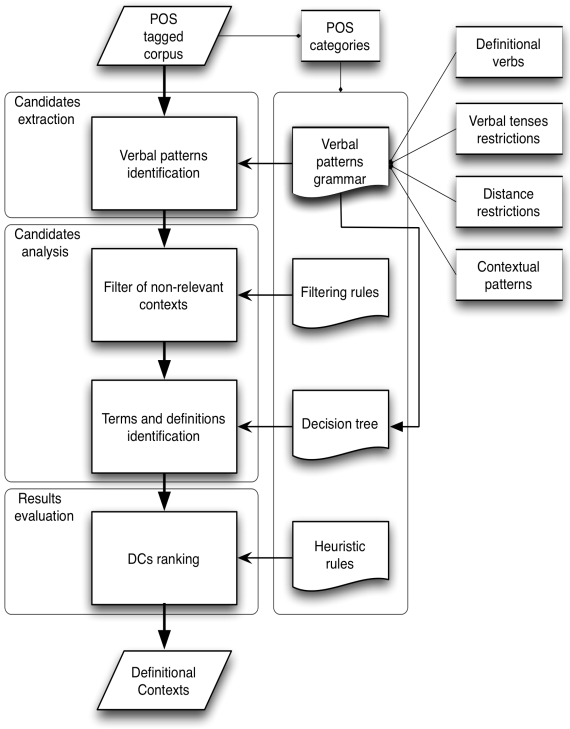

Fig. 1. System architecture.

4.1 Candidates extraction

The ECODE was developed taking the IULA's Technical Corpus from the Institut Universitari de Lingüística Aplicada (UPF) as starting point. This corpus consists of specialised documents in the fields of Law, Genome, Economy, Environment, Medicine, Informatics and General Language. First, we manually developed a grammar of verbal patterns for Spanish. We identified 29 verbs related to four different types of definitions: analytical, extensional, functional and synonimical. The whole set of verbal patterns is shown in table 1.

Table 1. Definitional Verbal patterns

Analytical verbal patterns
ser + artículo (to be + article)
consistir en (to consist in)
caracterizar como/por (to characterize as/for)
concebir como (to conceibe as)
considerar como (to consider as)
describir como (to describe as)
comprender como (to understand as)
definir como (to define as)
entender como (to understand as)
conocer como (to known as)
denominar como/∅ (to denominate as/∅)
llamar como/∅ (to call as/∅)
nombrar como/∅ (to name as/∅)
Extensional verbal patterns
comprender (to comprehend)
contener (to contain)
incluir (to include)
integrar (to integrate)
constar de (to comprise of)
contar de/con (to count of/with)
consistir de/en (to consist of/in)
formar de/por (to form of/by)
componer de/por (to compose of/by)
constituir de/por (to constitute of/by)
Functional verbal patterns
permitir (to allow)
encargar de (to undertake of)
funcionar como/para (to function as/for)
ocupar como/para (to occupy as/for)
servir como/en/para (to serve as/in/for)
usar como/en/para (to use as/in/for)
emplear como/en/para (to employ as/in/for)
utilizar como/en/para (to utilise as/in/for)
Synonimical verbal patterns
conocer también (to known also)
denominar (to denominate also)
llamar (to call also)
nombrar (to name also)

From the table above, we can see different verbs associated to different types of definitions. In some cases, the verbs can occur together with different grammatical particles and can be associated with more than one type of definition, such as the verb *denominar* (to denominate), which can occur in analytical or synonimical DCs with the nexus *como* (as) or *también* (also), respectively.

The verbal patterns were searched for taking into account the next constraints:

Verbal forms: infinitive, participle and conjugate forms.

Verbal tenses: present and past for verbs without nexus, any verbal tense for verbs with nexus.

Person: 3rd person singular and plural for verbs without nexus, any for verbs with nexus.

Distance: each nexus was searched for within a distance of 15 possible words.

With these restrictions, the system obtains a set of DC candidates that are next annotated with *contextual tags*. These simple tags function as borders in the next automatic processes. For each occurrence, the definitional verbal pattern was annotated with "<dvp></dvp>"; everything after the pattern with "<left></left>"; everything before the pattern with "<right></right>"; and finally, in those cases where the verbal pattern includes a nexus, like the adverb *como* (as), everything between the verbal pattern and the nexus was annotated with <nexus></nexus>. Here is an example of a DC annotated with contextual tags:

<left>El metabolismo</left> <dvp>puede definirse </dvp> <nexus>en términos generales como</nexus> <right>la suma de todos los procesos químicos (y físicos) implicados.</right>

4.2 Candidates analysis

Once the DCs were extracted and annotated with definitional verbal patterns they were analysed with the purpose of filtering non-relevant contexts. We applied this step based on the fact that definitional patterns are used not only in definitional sentences but also in a wider range of sentences. In the case of verbal patterns, some verbs tend to have a higher metalinguistic meaning than others. That is the case of *definir* (to define) or *denominar* (to denominate), vs. *concebir* (to conceive) or *identificar* (to identify), where the last two are used in different contexts. Moreover, verbs having a high metalinguistic meaning are not used only for defining terms.

To develop this process, a manual analysis was carried out to determine the type of grammatical particles or syntactic sequences occurring in those cases where a DVP was not used to define a term.

These particles and syntactic sequences were found in some specific positions, for example: negation particles such as *no* (not) or *tampoco* (either) were found in the first

position before or after the DVP; adverbs like *tan* (so), *poco* (few) as well as sequences like *poco más* (not more than) were found between the definitional verb and the nexus *como*; also, syntactic sequences such as adjective + verb were found in the first position after the definitional verb.

Thus, taking this and other frequently combinations into consideration as well as the contextual tags previously annotated, the systems filters contexts as shown in the following examples:

Rule: **NO <left>**

<left>En segundo lugar, tras el tratamiento eficaz de los cambios patológicos en un órgano pueden surgir problemas inesperados en tejidos que previamente **no </left>** <dvp>se identificaron</dvp> <nexus> como </nexus> <right> implicados clínicamente, ya que los pacientes no sobreviven lo suficiente.</right>

Rule: **<nexus> CONJUGATED VERB**

<left>Ciertamente esta observación tiene una mayor fuerza cuando el número de categorías </left> <dvp> definidas</dvp> **<nexus> es** pequeño como</nexus> <der>en nuestro análisis.</der>.

Once the non-relevant contexts were filtered, the next process was the identification of terms and definitions in the DC candidates. Depending on each DVP, the terms and definitions may appear in some specific positions in Spanish DCs. For example, in DCs containing the verb *definir* (to define), the term may occur in left, nexus or right position (T *se define como* D; *se define* T *como* D; *se define como* T D), while in DCs containing the verb *significar* (to signify), terms may appear only in left position (T *significa* D). Therefore, in this phase the automatic process is highly related to deciding the positions in which the constituent elements could appear.

We decided to use a decision tree to solve this problem, i.e., to detect by means of logic inferences the probable positions of terms, definitions and pragmatic patterns. We established some simple regular expressions to represent each constituent element[1]:

T = BRD (Det) + N + Adj. {0,2} .* BRD

PPR = BRD (sign) (Prep | Adv) .* (sign) BRD

D = BRD (Det) + N

As in the filtering process, the contextual tags function as borders to demarcate decision tree's instructions. In addition, each regular expression could function as a border. At the first level, the branches of the tree correspond to the different positions in which constituent elements may occur (left, nexus or right). At the second level, the branches correspond to the regular expressions of each DC element. The nodes (branches conjunctions) correspond to decisions taken from the attributes of each branch and are also horizontally related by *If* or *If Not* inferences, and vertically through *Then* inferences. Finally, the leaves correspond to the assigned position of each constituent element.

Hence, figure 2 shows an example of the decision tree inferences needed to identify constituent elements[2] in left position:

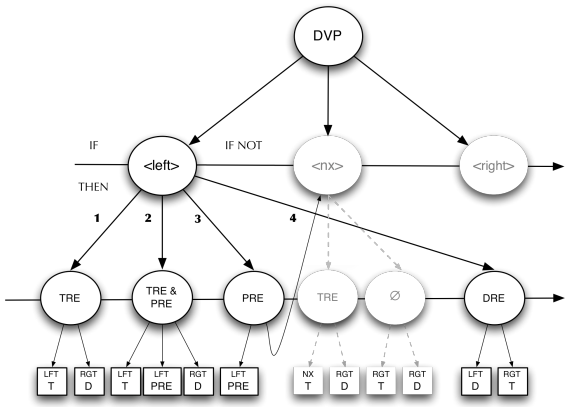

Fig. 2. Example of the identification of DCs elements.

This tree should be interpreted as follows: Given a series of DVPs occurrences:

1. *If* left position corresponds *only* to a term regular expression, *then*:

<left> = term | <right> = definition.

If Not:

2. *If* left position corresponds to a term regular expression and a pragmatic pattern regular expression, *then*:

<left> = term & pragmatic pattern | <right> = definition.

If Not:

3. *If* left position *only* corresponds to a pragmatic pattern regular expression, *then*[3]:

<left> = pragmatic pattern | *If* nexus corresponds *only* to a term regular expression, *then* <nexus> = term & <right> = definition; *If Not* <right> = term & definition.

4. *If* left position corresponds only to a definition regular expression, *then*:

[1] Where: Det= determiner, N= name, Adj= adjective, Prep= preposition, Adv= adverb, BRD= border and ".*"= any word or group of words.

[2] TRE = term regular expression | PRE = pragmatic pattern regular expression | DRE = definition regular expression.

[3] In some cases the tree must resort to other position inferences to find terms and definitions.

<left> = definition | <right> = term.

To exemplify this we can observe the next context:

"<left>En sus comienzos</left> <dvp>se definió</dvp> <nexus>la psicología como </nexus><right>"la descripción y la explicación de los estados de conciencia" (Ladd, 1887).</right>"

Once the DVP was identified as a CDVP – *definir como* (to define as) – the tree infers that left position:

1. Does not correspond only to a TRE.

2. Does not correspond to a TRE and a PRE.

3. It corresponds only to a PRE.

Then: left position is a pragmatic pattern (*En sus comienzos*). To identify the term and definition the tree goes to nexus's inferences and finds that:

1. It does correspond only to a TRE.

Then: nexus's position corresponds to the term (*la psicología*) and right's position corresponds to the definition ("la descripción y la explicación de los estados de conciencia […]").

As a result, the processed context was reorganised into terminological entries, as in the following example:

Table 2. Example of ECODE results

TERM	Psicología
DEFINITION	"la descripción y la explicación de los estados de conciencia" (Ladd, 1887)
PRAGMATIC PATTERN	En sus comienzos
VERBAL PATTERN	se definió como

4.3 Evaluation of results

In order to complement the system's processes described above, we decided to include an automatic ranking of the results. This automatic evaluation aims to identify those contexts with more prototypical structures of terms and definitions as well as structures reinforced by typographic markers.

Here, the input consists of candidates that were classified by the system as DCs, and a set of heuristic rules that analyse the syntactic structure of the elements automatically classified as term or definition is used to perform the ranking. Firstly, the ranking process assigns a numeric value to each identified term and definition of the candidates. Secondly, it combines those numeric values to generate a global value for each candidate.

Some of the heuristic rules can be seen in the next table:

Table 3. Example of ranking rules

Term = 1	<t>quotation marks .* quotation marks</t>
Term = 3	<t>.* pronoun .*</t>
Def = 1	<d>.* *that* .*</d>
Def = 3	<d>demostrative pronoun</d>

From the table above we can observe different rules that assign different values to the structure of terms and definitions. Value 1 means the best result, while 3 means the worst; candidates that do not follow any of the rules are assigned the value 2 by default. In the case of term's structures, the value 1 is assigned to those structures that are present between quotation marks, while a value of 3 is assigned to those candidates where the term structure consists of a pronoun, which could indicate a possible anaphoric reference. In the case of definition's rules, the value 1 is assigned to those structures where a relative clause is introduced after the pronoun *que* (that), which can be a prototypical structure in analytical definitions, while a value of 3 is given to the cases that consist only of a demonstrative pronoun. In the next table we illustrate some examples of each case:

Table 4. Example of ranking results

Term1	<t>«introns»</t>
Term3	<t>Este cloroplasto</t>
Definition1	<t>la mutación rutabaga</t> <dvp>es </dvp> <d>una mutación errónea que destruye a la adenilciclasa, interrumpiendo la síntesis del AMPc</d> .
Definition3	<d>Esto</d> <dvp>se conoce <nx> como</nx></dvp> <t>mutación</t>.

In the next sections we will describe our methodology for the system evaluation.

5. Evaluation

To develop the evaluation procedure we also used the IULA's technical corpus in Spanish. Taking into account that our system aims to identify DCs by searching for instances of definitional verbal patterns, we decided to set up a sub-corpus containing occurrences of the lemmas of the verbs from our grammar of verbal patterns. We searched for the first 250 occurrences of each verbal pattern (or all of the occurrences when they where less than 250), which produced a sub-corpus of 5809 sentences. Each one of those sentences was manually classified as DC or Non-Relevant, and was used as the input to the system to perform the evaluation.

We used precision & recall to evaluate the system performance. In this case, precision is the total number of

DCs automatically extracted, over the total number of candidates the system automatically identified as DCs, while recall is the number of DCs automatically extracted, over the total number of DCs presented in the evaluation sub-corpus.

The precision & recall results can be found in the next table:

Table 5. Precision & Recall results

P	R
0.53	0.79

It can be seen that almost the 80% of the total number of DCs was automatically extracted, while less than the 50% percent of the candidates was identified as noise, i.e., contexts that the system considers to be DCs but where manually tagged as Non-Relevant. In the case of recall, the system did not identify any candidates that were manually considered to be DCs.

In order to obtain a more specific scenario of the system's performance, we decided to apply an evaluation procedure for each kind of verbal patterns. For this purpose, we only considered those contexts containing one definitional verbal pattern. In this case, the sub-corpus consists of 4799 occurrences and the results are shown in the following table:

Table 6. Precision & Recall of definitional verbal patterns

Type	P	R
Analytical	0.58	0.83
Extensional	0.48	0.77
Functional	0.45	0.83
Synonimical	0.76	0.85

In general terms, it can be seen that the best results were obtained for synonimical patterns, while the lower values were obtained for the recall of the extensional patterns, and the precision for the functional patterns. These may be due to the fact that extensional patterns include verbs that can be used in a wider range of sentences and not only to introduce definitional information. Synonimical patterns, on the other hand, include verbs such as *conocer* (to known), *denominar* (to denominate), *llamar* (to call) and *nombrar* (to call) which, in conjunction with the particle *también* (also) seems to be more reliable for the recovering of definitional information. Analytical patterns show that some of the verbal forms can introduce a wider range of sentences that are considered to be noise. The same situation applies for the functional patterns.

6. Conclusions

We have presented a process of developing a definitional knowledge extraction system. This system aims at the simplification of the terminological practice related to the search for definitions of terms in specialised texts.

The methodology we have presented includes the searching for definitional patterns, the filtering of non-relevant contexts and the identification of DCs constituent elements: terms, definitions, and pragmatic patterns.

Up to now we have only worked with definitional verbs but we know that there is still further work to be done, which includes:

1. To explore other types of definitional patterns (mainly typographical patterns and reformulation markers) that are capable of recovering definitional contexts.
2. To improve the rules for the filtering process of non-relevant contexts, as well as to improve the algorithm for the automatic identification of constituent elements.
3. To improve the ranking algorithm.

Acknowledgments

This research was made possible by the financial support of the Consejo Nacional de Ciencia y Tecnología, Mexico and DGAPA-UNAM. The authors wish to thank the reviewers for its comments and suggestions, which helped to improve this paper.

7. References

[1] J. Pearson. Terms in Context. John Benjamin's, Amsterdam, 1998.

[2] I. Meyer. "Extracting Knowledge-rich Contexts for Terminography". In Recent Advances in Computational Terminology. D. Bourigault, C. Jacquemin and M.C. L'Homme (eds.).. John Benjamin's, Amsterdam, 2001. 278-302.

[3] J. Klavans and S. Muresan. "Evaluation of the DEFINDER System for Fully Automatic Glossary Construction". In Proceedings of the American Medical Informatics Association Symposium. ACM Press, New York, 2001. 252-262

[4] A. Storrer and S Wellinghoff. "Automated Detection and Annotation of Term Definitions in German Text Corpora". In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genève, 2006. 2373-2376.

[5] A. Pinto and D. Oliveira. Extracção de Definições no Corpógrafo [on line]. University of Porto, 2004. http://www.linguateca.pt/documentos/OliveiraPintoOut2004.pdf

[6] R. Alarcón, C. Bach and G. Sierra. "Extracción de contextos definitorios en corpus especializados: Hacia una elaboración

de una herramienta de ayuda terminográfica". Revista Española de Lingüística. Madrid, 2008. 247-278.

[7] V. Malaisé, P. Zweigenbaum and B. Bachimont. "Mining Definitional contexts to Help Structuring Differential Ontologies". Terminology 11 (1), 2005. 21-53.

[8] S. Walter and M. Pinkal. "Automatic Extraction of Definitions from German Court Decisions". In Proceedings of the Workshop on Information Extraction Beyond the Document. 21st International Conference on Computational Linguistics (COLING'2006). Sydney, 2006. 20–28.

[9] P. Monachesi. "The LT4eL Project: Overview" [on line]. University of Utrecht. 2007.

www.lt4el.eu/content/files/ws_prague/lt4el-prague.pdf

[10] H. Saggion. "Identifying Definitions in Text Collections for Question Answering". In *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC2004*. Lisboan, 2004. 1927-1930.

[11] R. Navigli and P. Velardi. "GlossExtractor: A Web Application to Automatically Create a Domain Glossary". In Lecture Notes in Computer Science 4733, 2007. 339-349

[12] F. Leu, and C. Ko. "An Automated Term Definition Extraction using the Web Corpus in Chinese Language". In Proceedings of the Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'07), 2007. 435-440.