ECODE: A Pattern Based Approach for Definitional Knowledge Extraction

Rodrigo Alarcón Martínez
Gerardo Sierra Martínez
Universidad Nacional Autónoma de México

Carme Bach Martorell
Universitat Pompeu Fabra

In this paper we present a pattern-based approach to the automatic extraction of definitional knowledge from specialised Spanish texts. Our methodology is based on the search of definitional verbal patterns to extract definitional contexts related to different kinds of definitions: analytic, extensional, functional and synonymic. This system could be a helpful tool in the process of elaborating specialised dictionaries, glossaries and ontologies.

1. Introduction

Mining definitional knowledge is becoming a growing interest for both Terminography and Natural Language Processing fields. Some efforts have been done in order to describe the process of defining terms in specialised texts (Pearson 1998, Meyer 2001), and have stated the premise that some recurrent lexical patterns are used in contexts where information to define a term is given. These types of contexts are defined as *definitional contexts* (DC) and are minimally composed by a term (T) and a definition (D), both related by a definitional pattern (DP) and occasionally modified by pragmatic patterns (PP) that state some relevant information about the scope and operation of the term in the context it appears.

Based on this idea, in this paper we will describe the methodology of developing the ECODE[1] system, an approach capable of extract definitional contexts from specialised corpora. Such methodology includes the extraction of definitional pattern's occurrences, the filtering of non-relevant contexts, and the identification of DCs constitutive elements, i.e., terms and definitions. This methodology is been developing for Spanish language and it could be seen as a helpful process in the elaboration of ontologies, lexical knowledge databases, glossaries or specialised dictionaries.

2. State of the art

The study of automatic extraction of definitional knowledge has been approached from both theoretical-descriptive and applied perspectives. One of the theoretical-descriptive works is Pearson's (1998), in which the behavior of the contexts where terms appear is described. Pearson describes that authors usually employ typographic patterns to visually bring out the presence of terms and/or definitions, as well as lexical and metalinguistic patterns to connect DCs elements by means of syntactic structures. This idea has been reinforced by Meyer (2001), who states that definitional patterns can also provide keys that allow the identification of the kind of definition present in DCs, which is a helpful task in the elaboration of ontologies. Other interesting theoretical-descriptive works can be found in Bach (2005) and Sierra et al. (2003).

Applied investigations leave from theoretical-descriptive studies with the objective of elaborate methodologies for the automatic extractions of DCs. Some of those applied investigations are the extraction of definitions in medical texts (Klavans & Muresan 2001), the extraction of

[1] "Extractor de Contextos Definitorios" (Definitional Contexts Extractor).

metalinguistic information (Rodríguez 2004), and the automatic elaboration of ontologies (Malaisé 2005). In general words, those studies employ definitional patterns as a common start point for the extraction of knowledge about terms. Our own related previous work on the linguistic description and analysis of DC contexts as well as their place in the task of elaboration a definitional knowledge extraction tool, can be found in Alarcón & Sierra (2006).

3. Methodology

The main purpose of a system for the automatic extraction of definitional knowledge would be to simplify the search of relevant information about terms, by means of searching the instances of definitional patterns. In this sense, a system that only retrieves those occurrences of definitional patterns would be for sure a useful system in different terminography tasks. However, a manual analysis of the occurrences would be necessary to detect those cases that do not provide definitional information about the terms, being their identification still a cost effort task. Therefore, we propose a methodology that includes not only the extraction of occurrences of definitional patterns, but also a filtering of non-relevant contexts (i.e., non definitional contexts) and the automatic identification of the possible constitutive elements of a DC: terms, definitions and pragmatic patterns.

In this study we took as reference the IULA's Technical Corpus, a corpus developed by the Institut Universitari de Lingüística Aplicada (IULA, UPF). We automatically search for instances of *definitional verbal patterns* (DVPs) which include *simple definitional patterns* (SDVP) like *concebir* (to conceive), *definir* (to define), *entender* (to understand), and *compound definitional patterns* (CDVP) which includes constructions like *consistir de* (to consist of), *consistir en* (to consist in), *denominar también* (also denominated), *usar como* (to use as)[2].

The approach we have been working on is based on the search of DVPs related to four different kinds of definitions: analytic, functional, extensional and synonymic. In figure 1 we represent the overall architecture of the system.

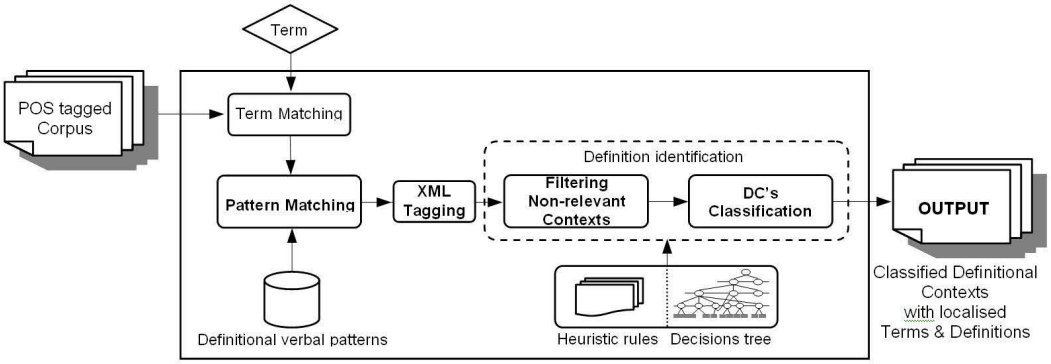

Figure 1. Overall ECODE architecture

The methodology uses a POS tagged corpus and takes a term as input. All the term occurrences in the corpus are extracted and become the instances for matching DVPs. The matched occurrences are tagged with XML, highlighting terms and DVPs. Finally they pass through a process of definition identification which includes a filtering of non-relevant contexts (i.e. those contexts where no definition is given), and a classification of DCs according to the type of definition they contain.

XML tagging is a simple process to annotate term's occurrences with DVPs. For each occurrence, the definitional verbal pattern were 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), for

[2] The complete set of definitional verbal patterns we have worked on can be seen on Table 1.

example in se define *como* (is defined as), everything between the verbal pattern and the nexus were annotated with <nexus></nexus>. Here is an example of a DC with XML tags.

> <LEFT>El metabolismo</LEFT> <DVP>puede definir se</DVP> <NEXUS>en términos generales como</NEXUS> <RIGHT>la suma de todos los procesos químicos (y físicos) implicados.</RIGHT>

For the filtering of non-relevant contexts we previously made an analysis to determine which kind of grammatical particles or syntactic sequences could appear in those cases when a DVP is not used to define a term. Those particles and sequences were found in some specific positions, for example: some negation particles like *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 like *adjective* + *verb* were found in the first position after the definitional verb.

To perform the definition identification we use a decisions tree based on heuristics to detect by means of logic inferences the position of the definition and pragmatic patterns if they appear. In Spanish's DCs, and depending on each DVP, the terms and definitions can appear in some specific positions. For example, in DCs with the verb *definir* (to define), the term could appear in left, nexus or right position (T *se define como* D; *se define* T *como* D; *se define como* T D), while in DCs with the verb *significar* (to signify), terms can appear only in left position (T *significa* D). The tree uses the next set of regular expressions to represent each constitutive element[3]:

TRE	=	query term
PPRE	=	BRD (sign) (Prep | Adv) .* (sign) BRD
DRE	=	BRD Det. + N .* BRD

A representation of decisions tree could be found in the next figure.

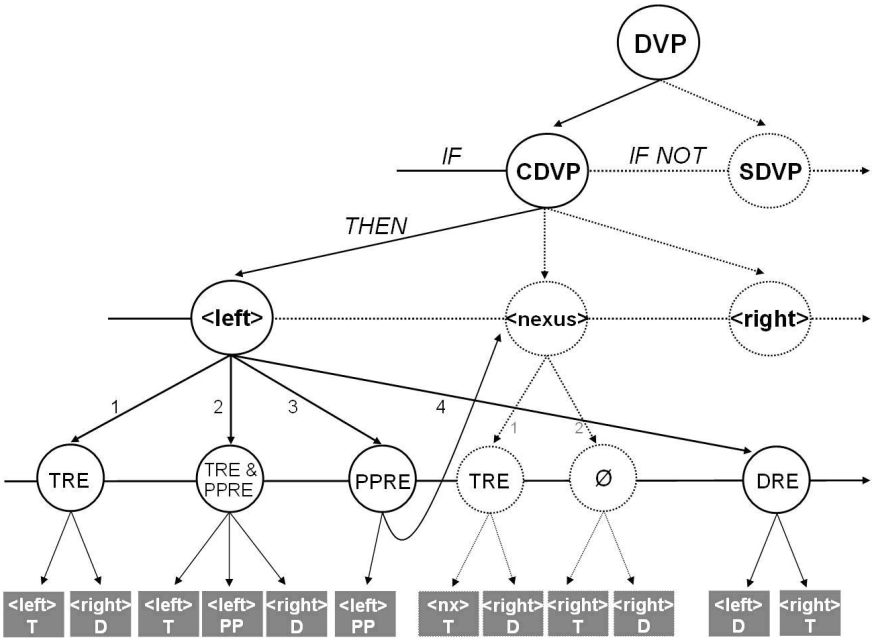

Figure 2. Decisions tree

[3] Where: TRE = Term Regular Expression, PPRE = Pragmatic Pattern Regular Expression, DRE = Definition Regular Expression, Det= determiner, N= name, Adj= adjective, Prep= preposition, Adv= adverb, BRD= border, and ".*"= any word or group of words.

In a first level, the branches of the tree are the different positions in which constitutive elements can appear (left, nexus or right). In a second level, the branches are the regular expressions of each DC element, where the borders could be the XML tags as well as the regular expressions of other constitutive elements. The nodes (branches conjunctions) corresponds to decisions taken from the attributes of each branch and also are horizontally related by *If* or *If Not* inferences, and vertically through *Then* inferences. Finally, the leaves are the assigned position for a constitutive element.

To exemplify we present a result for the term "genotipo". Given a context like the next one:

> Cualquier individuo de la población puede poseer uno de los tres genotipos posibles (recuérdese que en el capítulo 3 se definió el genotipo como la constitución genética de un individuo en un locus).

The algorithm identifies it as a DC and tags both the definition type in function of the DVP and each constitutive element, given as result the next XML tagged context:

> <DC type= "analytical">Cualquier individuo de la población puede poseer uno de los tres genotipos posibles (recuérdese que en el capítulo 3 <DVP><VP>se definió</VP> el <TERM> genotipo</TERM> <NX>como</NX></DVP> <DEFINITION>la constitución genética de un individuo en un locus <DEFINITION>).</DC>

The system organises the annotated contexts into a sort of basic terminographical entries that include the specification of the type of definition, the associated verbal pattern and the number of DCs and Non relevant contexts automatically extracted. These basic terminological entries (figure 3) show the term and its correspondent definition, the complete definitional context and the source corpus code where they were founded. More details and examples can be found at *http://brangaene.upf.es/ecode*.

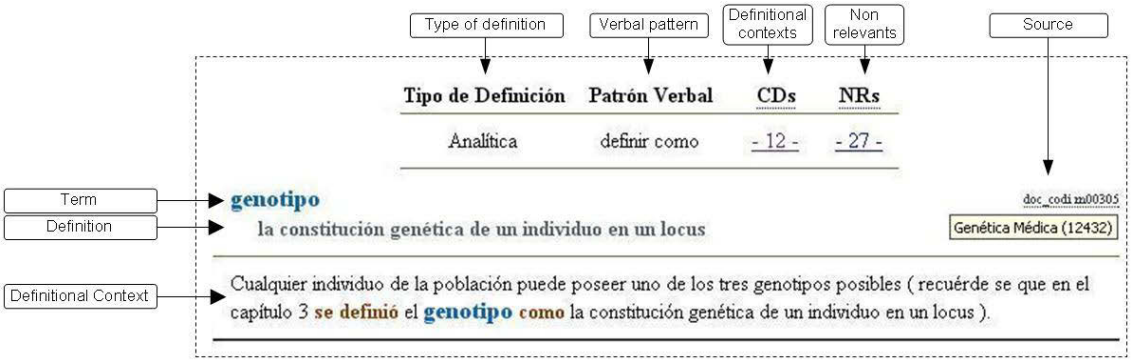

Figure 3. Example of the results

4. Evaluation

We evaluate the extraction of DVPs and the filtering of no relevant contexts using Precision & Recall. In general words, Precision measures how many information extracted is *relevant*, while Recall measures how many *relevant* information was extracted from the input. In our case, Precision consists of the total number of filtered DCs automatically extracted over the total number of contexts automatically extracted; while Recall consists of the total number of filtered DCs automatically extracted, over the total number of *non filtered* DCs automatically extracted[4]. Results can be seen in table 1 (a number close to 1 indicates a better result).

[4] We use the number of non filtered DCs extracted for measuring Recall due to our intention of evaluating the filtering process.

Verbal pattern		P	R
Concebir (como)	To conceive (as)	0.67	0.98
Definir (como)	To define (as)	0.84	0.99
Entender (como)	To understand (as)	0.34	0.94
Identificar (como)	To identify (as)	0.31	0.90
Consistir de	To consist of	0.62	1
Consistir en	To consist in	0.60	1
Constar de	To comprise	0.94	0.99
Denominar también	Also denominated	1	0.87
Llamar también	Also called	0.90	1
Servir para	To serve for	0.55	1
Significar	To signify	0.29	0.98
Usar como	To use as	0.41	0.95
Usar para	To use for	0.67	1
Utilizar como	To utilise as	0.45	0.92
Utilizar para	To utilise for	0.53	1

Table 1. Results of Precision & Recall

In the case of Precision, there is a divergence on verbs that usually appear in metalinguistic sentences. The best results were obtained with verbs like *denominar* (to denominate) or *definir* (to define), while verbs like *entender* (to understand) or *significar* (to signify), which can be used in a wide assortment of sentences, (not necessarily DCs), recover low Precision values. In the case of Recall, higher results indicate that fewer valid DCs were filtered as non-relevant contexts. The wrong filtering of DCs as non-relevant contexts is related to the non-filtering rules, but also in some cases a wrong classification was due to a POS tagging errors in the input corpus.

The challenge we face to improve the results is directly related to the elimination of noise. We have noticed that the more precise the DVP is, the better results (in terms of less noise) can be obtained. Nevertheless, a specification of DVPs means a probable lost of recall. Although, a revision of filtering rules must be done in order to improve the non-relevant contexts identification and avoid the cases when some DC were incorrect filtered.

5. Conclusions

In this paper we have presented the methodology that constitutes a definitional knowledge extraction system. The aim of this approach is the simplification of the terminographical practices related to the search of term's definitions in specialised texts. The methodology we have presented includes the search of definitional verbal patterns, the filtering of non-relevant contexts and the identification of constitutive elements, i.e., terms, definitions and pragmatic patterns.

Although we have worked with definitional verbs, there is still a lot of work to be done in order to improve the system we have presented. We are currently working on the optimisation of the filtering rules to perform a better identification of DCs. It is necessary to continue with the formal description of all patterns that constitute a DC and to observe the possible role that these other patterns play for establishing new patterns for the automatic extraction of DCs.

References

Alarcón, R.; Sierra, G. (2006). "Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios". In Hernández, A.; Zechinelli, J. L. (eds.). *Avances en la Ciencia de la Computación, VII Encuentro Nacional de Ciencias de la Computación*. San Luís Potosí: MSCC. 242-247.

Bach, C. (2005). "Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado". *Debate Terminológico* 1 [on-line]. [Acces date 26 March 2008]. *http://www.riterm.net/revista/n_1/bach.pdf*.

Klavans, J.; Muresan, S. (2001). "Evaluation of the DEFINDER system for fully automatic glossary construction". In *Proceedings of the American Medical Informatics Association Symposium*. New York: ACM Press. 252-262.

Malaisé, V. (2005). Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels. Ph.D. Paris: Université Paris 7 – Denis Diderot.

Meyer, I. (2001). "Extracting Knowledge-rich contexts for Terminography". In Bourigault, D.; Jacquemin, C.; l'Homme, M. C. (eds.). *Recent advances in Computational Terminology*. Amsterdam: John Benjamins. 278-302.

Pearson, J. (1998). *Terms in context*. Amsterdam: John Benjamins.

Rodríguez, C. (2004). "Metalinguistic Information Extraction for Terminology". In Ananiadou, S.; Zweigenbaum, P. (eds.). *3rd International Workshop on Computational Terminology*. Geneva: Coling. 15-22.

Sierra, G. et al. (2003). "Definitional Contexts Extraction from Specialised Texts". In Lewandowska, B. (ed.). *Practical Applications in Language and Computers*. Frankfurt am Main: Peter Lang. 21-31.