

Developing a Definitional Knowledge Extraction System

Rodrigo Alarcón¹, Gerardo Sierra^{1,2}, Carme Bach²

¹ Grupo de Ingeniería Lingüística, Instituto de Ingeniería, Universidad Nacional Autónoma de México

² Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra

{ralarconm, gsierram}@ii.unam.mx, carme.bach@upf.edu

Abstract

One of the main goals of terminological work is the identification of knowledge about terms in specialised texts. In order to compile dictionaries, glossaries or ontologies, terminographers use to search definitions about the terms they intend to define. The search of definitions can be done in specialised corpus, where they usually appear in definitional contexts, i.e. text fragments where an author explicitly defines a term. In this paper we present a research focused on the automatic extraction of those definitional contexts. Our methodology includes three different processes: the extraction of definitional patterns, the automatic filtering of non definitional contexts, and the automatic identification of constitutive elements, i.e., terms and definitions.

1. Introduction

A common need in terminological work is the extraction of knowledge about terms in specialised texts. Some efforts in the field of NLP have been done in order to develop tools that help in this need, such as corpora, where a large quantity of technical documents are digitally stored, and term extraction systems, which automatically identify relevant terms in corpora.

Nowadays there is a growing interest on developing systems for the automatic extraction of useful information to describe the meaning of terms. This information commonly appears in structures called *definitional contexts* (DCs), which are structured by a series of lexical and metalinguistic patterns that can be automatically recognised (Pearson, 1998; Meyer, 2001). Following this idea, our work is focused on developing a system for the automatic extraction of definitional contexts on Spanish language specialised texts. Such system 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 system is been developing for Spanish language and it will be helpful in the elaboration of ontologies, databases of lexical knowledge, glossaries or specialised dictionaries.

In this paper we will describe the structure of DCs; we will make a short review of related works; we will present the methodology we have followed for the automatic extraction of DCs, in addition with a methodology's evaluation; and finally we will describe the future work.

2. Definitional Contexts

A definitional context is a textual fragment from a specialised text where a definition of a term is given. It is basically structured by a term (T) and its definition (D), and both elements are connected by typographic or syntactic patterns. Mainly, 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)–. Besides, DCs can include pragmatic patterns (PP), which provide conditions for the use of the term or clarify its meaning,

like *en términos generales* (in general terms) or *en este sentido* (in this sense). The next is an example of a definitional context:

→ 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), and the general sense of the contexts is modify by a pragmatic pattern *desde un punto de vista práctico* (from a practical point of view).

2.1. 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 first theoretical-descriptive works is Pearson's (1998), in which the behaviour of the contexts where terms appear is described. Pearson mentions that, when authors define a term, they 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 theoretical-descriptive works can be found in Feliu (2004) and Bach (2001 & 2005).

Applied investigations, on the other hand, leave from theoretical-descriptive studies with the objective of elaborate methodologies for the automatic extractions of DCs, more specifically for the extraction of definitions in medical texts (Klavans & Muresan, 2001), for the extraction of metalinguistic information (Rodríguez, 2004), and for 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. For developing the methodology we present in this paper we leave from the analysis and integration of theoretical-descriptive and applied studies. In the next section we describe in detail our methodology.

3. Definitional Contexts Extraction

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

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

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.

3.1. Corpus

We took as reference the IULA's Technical Corpus and its search engine bwanaNet¹. This corpus was developed by the Instituto Universitario de Lingüística Aplicada (IULA, UPF). It is formed by specialised documents in Law, Genome, Economy, Environment, Medicine, Informatics, and General Language. It counts with a total of 1,011 documents (July 2006). For the experiments we describe here we use all the areas except General Language, and the number of treated documents was 959 with a total number of 11,569,729 words.

3.2. Extracting Definitional Patterns

For the experiments that we describe in this paper we searched for definitional verbal patterns (DVPs). We worked with 15 patterns which include *simple definitional verbal patterns* (SDVP) and *compound definitional verbal patterns* (CDVP). Patterns of the simple forms include only the definitional verb, while patterns of the compound forms include the definitional verb plus a grammatical particle such as a preposition or an adverb:

SVDP: *concebir* (to conceive), *definir* (to define), *entender* (to understand), *identificar* (to identify) and *significar* (to signify).

CVDP: *consistir de* (to consist of), *consistir en* (to consist in), *constar de* (to comprise), *denominar también* (also denominated), *llamar también* (also called), *servir para* (to serve for), *usar como* (to use as), *usar para* (to use for), *utilizar como* (to utilise as) and *utilizar para* (to utilise for).

Each pattern was searched in the Technical IULA's corpus through the *complex search* option, which allows users to obtain the occurrences with POS tags. We also delimitate the search to no more of 300 occurrences for each verbal pattern, using the random recovery option.

The verbal patterns were searched taking into account the next restrictions:

Verbal forms: infinitive, participle and conjugate forms.

Verbal tenses: present and past for the simple forms, any verbal time for the compounds forms.

Person: 3rd singular and plural for the simple forms, any for the compound forms.

The obtained occurrences were automatically annotated with *contextual tags*. The function of these simple tags is to work as borders in the next automatic process. 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), everything between the verbal pattern and the nexus were annotated with <nexus></nexus>.

Here is an example of a DC with contextual 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>
```

It is important to mention that from this contextual annotation process, all the automatic process was done with scripts in Perl. We choose this programming language mainly by its inherent effectiveness to process regular expressions.

3.3. Filtering non-relevant contexts

Once we have extracted and annotated the occurrences with DVPs, the next process was the filtering of nonrelevant contexts. We apply this step based on the fact that definitional patterns are not used only in definitional sentences. In the case of DVPs some verbs trend to have a high metalinguistic meaning rather 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 ones could be used in a wide variety of different sentences. Moreover, the verbs with a high metalinguistic meaning are not used only for defining terms.

In a previous work (Alarcón & Sierra 2006) an analysis was done in order 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.

Thus, considering this and other frequently combinations and helped by contextual tags previously annotated, we developed a script in order to filtering non-relevant contexts. The script could recognise contexts like 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>
```

¹ <http://bwananet.iula.upf.edu/indexes.htm>

3.4. Identifying DCs elements

Once the non-relevant contexts were filtered, the next process in the methodology is the identification of main terms, definitions and pragmatic patterns.

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). Therefore, in this phase the automatic process is highly related to deciding in which positions could appear the constitutive elements.

We decided to use a decision tree (Alarcón, 2006) 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 constitutive element²:

- T** = BRD (Det) + N + Adj. {0,2} .* BRD
- PP** = BRD (sign) (Prep | Adv) .* (sign) BRD
- D** = BRD Det. + N .* BRD

As well as in the filtering process, the contextual tags have functioned as borders to demarcate decision tree's instructions. In addition, each regular expression could function as a border.

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. 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. Hence, in figure 1 we present an example of the decision tree inferences to identify left constitutive elements³. This tree should be interpreted in the next way:

- Given a series of DVPs occurrences:

If verbal pattern = compound definitional verbal pattern, *then*:

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*⁴:

→ <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*:

<left> = definition | <right> = term

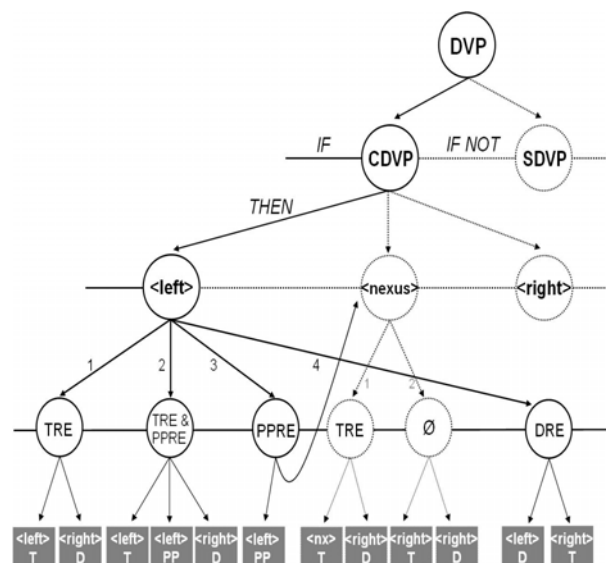


Figure 1: Identification of left position's elements

To exemplify 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 PPRE.
3. It does correspond only to a PPRE.

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 result, the processed context was reorganised into terminological entries as in the next example:

Term	psicología
Definition	"la descripción y la explicación de los estados de la conciencia" (Ladd, 1887).
Verbal pattern	se define como
Pragmatic pattern	En sus comienzos

Table 1: Example of constitutive elements identification

To conclude this part we have to mention that the algorithms we developed implement non complex regular expressions as well as simple logic inferences to find, analyse and organise definitional knowledge. Furthermore, the design of the algorithms allows the implementation in other languages by replacing the correspondent regular expressions as well as the logical inferences.

² Where: **Det**= determiner, **N**= name, **Adj**= adjective, **Prep**= preposition, **Adv**= adverb, **BRD**= border and ".*"= any word or group of words.

³ **TRE** = term regular expression | **PPRE** = pragmatic pattern regular expression | **DRE** = definition regular expression.

⁴ In some cases the tree must resort to other position inferences to find terms and definitions.

4. Evaluation

The evaluation of the methodology consists in two parts:

- 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.
- For the identification of constitutive elements, we manually assigned values that helped us to statistically evaluate the exactitude of the decisions tree.

4.1. Evaluation of DVPs extraction and non-relevant contexts filtering

We determine Precision & Recall by means of the following formulas:

P = the number of filtered DCs automatically extracted, over the number of contexts automatically extracted.

R = the number of filtered DCs automatically extracted, over the number of *non filtered* DCs automatically extracted⁵.

The results for each verbal pattern can be seen in table 2⁶. 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) recover low Precision values. Those verbs with lower results can be used in a wide assortment of sentences, (i.e., not necessarily definitional contexts), and they trend to recover a big quantity of noise.

In the case of Recall, low results indicate that valid DCs were filtered as non-relevant contexts. The wrong classification 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.

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 2: Results of Precision & Recall

⁵ We use the number of non filtered DCs extracted for measuring Recall, due to our intention of evaluating the filtering process.

⁶ A number close to 1 indicates a better result.

The challenge we face in this stage is directly related to the elimination of noise. We have noticed that the more precise the verbal pattern is, the better results (in terms of less noise) can be obtained. Nevertheless, a specification of verbal patterns 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 where incorrect filtered.

4.2. Evaluation of DCs elements identification

To evaluate the DCs elements identification, we manually assign the next values to each DC processed by the decisions tree:

3 for those contexts where the constitutive elements were correct classified;

2 for those contexts where the constitutive elements were correct classified, but some extra information were also classified (for example extra words or punctuation marks in term position);

1 for those contexts where the constitutive elements were *not* correct classified, (for example when terms were classified as definitions or vice versa).

Finally, with the symbol \emptyset we designate the contexts that the system could not classify.

In table 3 we present the results of the evaluation of DCs elements identification. The values are expressed as percentages, and the amount of all of them represent the total number of DCs founded with each verbal pattern.

From Dcs evaluation we highlight the following facts:

- The average percentage of the correct classified elements (group “3”) is over the 50 percent of the global classification. In these cases, the classified elements correspond exactly with a term or a definition.

- In a low percentage (group “2”), the classified elements include extra information or noise. Nevertheless, in these cases the elements where also good classified as in group “3”.

- The incorrect classification of terms and definitions (group “1”), as well as the unclassified elements (group “ \emptyset ”) correspond to a low percentage of the global classification.

Verbal pattern	3	2	1	\emptyset
Concebir (como)	68.57	15.71	11.42	04.28
Definir (como)	65.10	18.22	10.41	06.25
Entender (como)	54.16	20.83	08.33	16.66
Identificar (como)	51.72	05.17	34.48	08.62
Consistir de	60.00	0	20.00	20.00
Consistir en	60.81	8.10	15.54	15.54
Constar de	58.29	22.97	02.97	15.74
Denominar también	21.42	28.57	07.14	42.85
Llamar también	30.00	40.00	0	30.00
Servir para	53.78	27.27	0.007	18.18
Significar	41.26	44.44	03.17	11.11
Usar como	63.41	14.63	17.07	04.87
Usar para	36.26	32.96	04.39	26.37
Utilizar como	55.10	28.57	10.20	06.12
Utilizar para	51.51	19.69	10.60	18.18

Table 3: Evaluation of DCs elements identification

Since the purpose of this process was the identification of DCs elements, we can argue that results are generally satisfactory. However, there is a lot of work to do in order to improve the performance of decision's tree inferences. This work is related to the way the tree analyses the different DCs elements of each verbal pattern.

5. Conclusions and future work

In this paper we have presented the process of developing a definitional knowledge extraction system. The aim of this system is the simplification of the terminological practice related to the search of term's definitions in specialised texts.

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

At this moment we have worked with definitional verbs and we know that there is a lot of work to do, which basically consists of the following points:

- a) To explore other kind of definitional patterns (mainly typographical patterns and reformulation markers) that are capable to recover definitional contexts.
- b) To include those definitional patterns mentioned above in each step of the methodology.
- c) To improve the rules for the non-relevant contexts filtering process, as well as the algorithm for the automatic identification of constitutive elements process.

Acknowledgments

This research has been developed by the sponsorship of the Mexican National Council of Science and Technology (CONACYT), by grants 179210 and 46832- H, as well as the Macro Project *Tecnologías para la Universidad de la Información y la Computación*, UNAM.

We also acknowledge the help of Bertha Lecumberri in the translation of this paper.

References

- Alarcón, R. (2006). Extracción automática de contextos definitorios en corpus especializados. Propuesta para el desarrollo de un ECCODE (extractor de candidatos a contextos definitorios). Phd project thesis. Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.
- Alarcón, R. & Sierra, G. (2006). Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios. In *Avances en la Ciencia de la Computación, VII Encuentro Nacional de Ciencias de la Computación* (pp. 242-247). Hernández, A., Zechinelli, J.L. (eds). San Luís Potosí: MSCC.
- Bach, C. (2001). La equivalencia parafrástica en los textos especializados en vista a la detección de información paralela. In *La terminología científico-técnica: reconocimiento análisis y extracción de informacación formal y semántica* (pp. 217-226). Cabré, M.T., Feliu, J. (eds). Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.
- Bach, C. (2005). Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado. In *Debate Terminológico. Electronic Journal*. Num. 1. Ríterm.
- Feliu, J. (2004). Relaciones conceptuales i terminologia: anàlisi i proposta de detecció semiautomàtica. PhD thesis. Barcelona: Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.
- Klavans, J. & Muresan, S. (2001). Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proceedings of the American Medical Informatics Association Symposium* (pp. 252-262). New York: ACM Press.
- Malaisé, V. (2005). Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels. Phd thesis. Paris: UFR de Linguistique, Université Paris 7 – Denis Diderot.
- Meyer, I. (2001). Extracting Knowledge-rich contexts for Terminography. In *Recent advances in Computational Terminology* (pp. 278-302). Bourigault, D., Jacquemin, C., L'homme, M.C. (eds). Amsterdam: John Benjamin's.
- Pearson, J. (1998) *Terms in context*. Amsterdam: John Benjamin's.
- Rodríguez, C. (2004). Metalinguistic Information Extraction for Terminology. In *3rd International Workshop on Computational Terminology* (pp. 15-22). Ananiadou, S., Zweigenbaum, P. (eds). Geneva: Coling.
- Saggion, H. (2004). Identifying Definitions in Text Collections for Question Answering. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1927--1930). Lisbon: European Language Resources.