Pattern Construction for Extracting Domain Terminology

Yusney Marrero García
Agrarian University of Havana, Cuba
yusneym@unah.edu.cu

Paloma Moreda Pozo
University of Alicante, Spain
moreda@dlsi.ua.es

Rafael Muñoz Guillena
University of Alicante, Spain
rafael@dlsi.ua.es

Abstract

The extraction of domain terminology is a task that is increasingly used for different application processes of natural language such as the information recovery, the creation of specialized corpus, question-answering systems, the creation of ontologies and the automatic classification of documents. This task of the extraction of domain terminology is generally performed by generating patterns. In literature we could find that the patterns which are used to extract such terminology often change from one domain to another, it means the intervention of human experts to the generation and validation of these patterns. This article deals with a methodology for automatic obtaining patterns (Basic Patterns and Definitory Verbal Patterns) for extracting domain terminology and minimizing the manual work of the experts. The obtained methodology was evaluated in the computer science domain obtaining a 97 percent in the case of the values of the basic patterns and a 98 percent of the definitory verbal patterns. Then the methodology was tested in three other domains with similar results, Agricultural Engineering (a 96 percent of the basic patterns and a 97 percent of the definitory verbal patterns), Veterinary Medicine (98% of the basic pattern and the definitory verbal patterns) and Agronomy (96% of the basic pattern and the definitory verbal patterns), showing that methodology can be applied in any specialty curriculum documents.

1 Introduction

The extraction of terms that characterizes a document is a task of vital importance in the development of recovery systems and information extraction.

It is very important to get the patterns that characterize these terms for the proper functioning of such systems.

In the present systems of natural language processes, there is a tendency which minimizes the human labor, leaving the processing of the whole information to the system but the final validation made by experts remains irreplaceable in many cases.

Sometimes the obtained patterns change from one domain to another, so there are some methods to minimize the human intervention. It would be a great step forward for the work of such systems.

The research paper is organized as follows: after presenting the state of the art (Section 2), we present in Section 3 Pattern Generation Process, Selection of the corpus (Section 3.1), Definitory context (Section 3.2), Definitory verbal patterns (Section 3.3) and Our proposal (Section 3.4). Then, the processes of Evaluation – Analysis for the Computer Science domain (Section 4) and then the evaluation of the obtained methodology in other domains will be presented (Section 4.1) and the ending with conclusions and future work (section 5).

2 State of the Art

Obtaining patterns for the extraction systems are a task whose success will depend on the correct operation of the system that uses it, the values of recall and precision have a direct correspondence with the obtained information in mapping with patterns.

Several proposals have been introduced to try to solve this problem such as (Riloff, 1993) and (Soderland et al., 1995). In these proposals as well as in the presented methodology extraction patterns are generated and are based on annotated corpus of training. The process of annotation of a corpus is clearly easier than the creation of a pattern dictionary manually, although it is true that it requires a domain expert that conducts and supervises the labeling of the

Proceedings of Recent Advances in Natural Language Processing, pages 420–426,
Hissar, Bulgaria, Sep 7–9 2015.

corpus.

Other proposals have avoided the annotation process like (Riloff and Shoen, 1995) in this case an annotated corpus is not required but preclassified, that is to say that the texts which are received as input must have been previously classified and (Huffman, 1995) in which a user is allowed to identify entities of interest that may represent events of interest.

3 Pattern Generation Process

It is very important to identify the recurrent syntactic structures of the terms that characterize these texts so as to extract domain terminology in specialized texts automatically. These structures represent patterns that follow the terminology that characterizes this domain. (Saneifar et al., 2009), (Sierra et al., 2006).

So as to develop an automatic terminology extraction domain which is based on patterns, the correct identification of these structures is a key factor for its proper operation.

In specialized texts it is very common to find many of the terms that characterize the texts.

In this section and these subsections our methodology is presented. It deals with the automatic generation of patterns to extract domain terminology in Spanish as well as other elements needed to understand it. The proposed methodology is based on two sets of patterns, the Basic Patterns and Definitory Verbal Patterns, the latter are incorporated in the methodology with the aim of improving the obtained precision of values which is based on the idea that most of the terms are defined in domain texts, belonging to them, which are framed in defining contexts as proposed by (Alarcon et al., 2007).

Next, a description of the corpus selection process is presented and the reasons for their selection.

3.1 Selection of the Corpus.

The selection of the corpus to use is a difficult but important task, because it is going to get the language patterns of the terms that are going to be used for tests and evaluation processes.

As proposed by (Dubuc and Lauriston, 1997), so as to elect the corpus we must take into account that:

- The text must be representative. The document scanning object has to reflect the use of experts in a specialty field.
- The nature of the publication largely determines the importance of contexts it contains. Textbooks, manuals, monographs, are excellent sources that provide explicit information of concepts and terms. The analysis of random samples of texts in a publication may determine its usefulness for terminology research.
- We must pursue a minimum of presentation and reliability. In general, poorly written texts with many grammatical mistakes provide a little solid base of terminological analysis.

Following the recommendations of Dubuc and Lauriston, some documents have been selected as corpus (120 documents in Spanish) these documents deal with the subjects belonging to the Curriculum Base and Own of the study Plan "D" of the Computer Science career of the Agrarian University of Havana paying special attention to the texts of each curriculum that are generally representative, reviewed and approved by experts in each domain, they are variegated in different areas where each domain is composed by a continuous updating. Texts provide a very important content having a correct presentation and reliability due to the staff and the destination where they will be used.

3.2 Definitory Contexts.

In (Sierra, 2009), a study with different approaches to the concept of Context Definitory (CD) is made in terminology (De Bessé, 1991), (Auger, 1997), (Pearson, 1998) and (Meyer, 2001).

In (Alarcon et al., 2007), the term CD deals with any textual fragment of a specialized document where a term is defined. CDs are formed by a term (T) and a definition (D), which are connected by a defining pattern (PD). They may optionally include a pragmatic pattern (PP), that is to say, structures that provide conditions using this term or qualifying its meaning. Figure 1

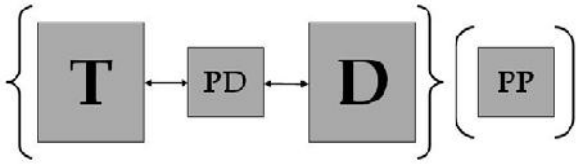

Figure 1: Structure of a defining context

3.3 Definitory Verbal Patterns.

(Alarcon, 2009) suggests that there are syntactic patterns that connect the term with its definition, if such connectors have a verb as the nucleus, then we have a Definitory Verbal Pattern (DVP).

In this sense we could find specialized texts with DVP.

Example 1:

Así, **se define** el *estándar XML* **como** el formato universal para documentos y datos estructurados en Internet y podemos explicar las características de su funcionamiento a través de 7 puntos importantes, tal y como la propia W3C recomienda.

Example 2:

cliente servidor: **Es una** tendencia de los actuales sistemas de operación que consiste en instrumentar la mayoría de las funciones en procesos usuarios, construyendo un "kernel" mínimo.

In the above examples we observe that the defining information is composed by the verbs *define* and *be*. Furthermore, the occurrence of the pronoun *se* to the verb *define*, and the adverb *como* to form the pattern *se define como*. In Example 2, we have the combination *es un*, a prototypical structure to define a term.

3.4 Our Proposal

In Figure 2 we present our methodology of automatic extraction of patterns where every step is described below.

1. Selection of the corpus belonging to the domain.

The first step of our methodology is to select the corpus we are going to use. This corpus should be divided into two parts; one part is used in the process of obtaining patterns and the remaining part in the evaluation process.

2. Semi-automatic annotation of terms belonging to the domain in question (Human expert validation)

For the labeling process we have constructed the **TermEt** tool which basically has two functions:

a) If you do not have a set of patterns already obtained, you have to show a view of the text and experts will be able to mark and write notes about the terms which belong to the domain.

b) If there is a set of patterns, the application allows their input showing the word or strings of words to be mapped with the previously introduced patterns, allowing the expert labelling or not the terms with the same tags.

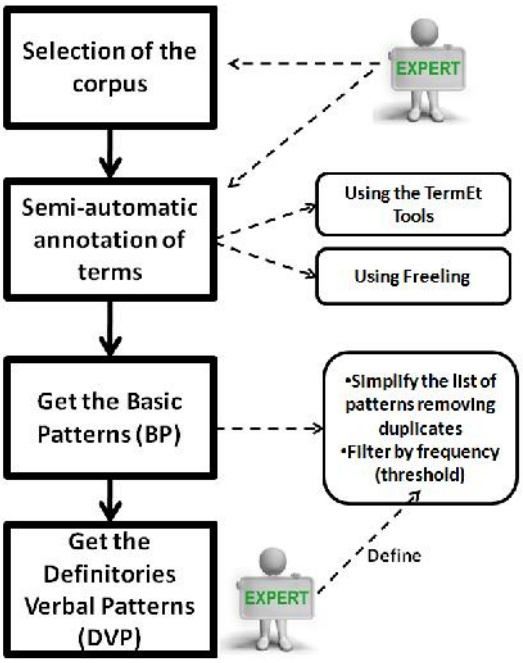

Figure 2: Our Methodology

In both cases, a morphological analysis is performed to the text using the Freeling[1] tool, and as an output an XML file is provided with the processed text and the terms which have been listed with their corresponding grammatical categories.

3. Get the basic patterns. This process involves the extraction of the label string obtained from a morphological analysis to the words that were annotated in the corpus as a term. Simplify the list of patterns removing duplicates and filter it through its frequency.

From the XML the obtained file as an output from the previous step and the list of strings for the terms which were included in the processed documents were extracted and a first set of possible patterns is obtained. Table 1 shows a fragment of the initial list of obtained patterns.

	Patterns
N	Noun
NJ	Noun+ Adjective
N	Noun
N	Noun
NPN	Noun+Preposition+Noun
NPNJ	Noun+Preposition+Noun+Adjective
N	Noun
NJCJ	Noun+Adjective+Conjuntion+Adjective

Table 1: Initial list of patterns

[1] http://nlp.lsi.upc.edu/freeling/

The number of obtained patterns may be very large, in order to simplify this pattern list we have to eliminate duplicated patterns and the frequency of each one is stored A filtering process is then performed, its frequency of appearance in the text is emphasized, experts can set a threshold and all patterns that its frequency in the text do not exceed this threshold will be deleted from the final list of patterns.

Table 2 shows the final list of resulting patterns after making the pro-filtering process considering the frequency of occurrence in each of the patterns.

This set of obtained patterns are called Basic Patterns (BP), and they represent the basic structures that follow the terms of a particular domain.

Patterns		**Frequency**
N	Noun	F1
NJ	Noun+ Adjective	F2
NPN	Noun+Preposition+Noun	F3

Table 2: Basic Patterns

As we can see if we use this set of obtained patterns we will surely obtain the terms that define that domain, but they are so basic that a lot of noise will be introduced affecting the precision values severely.

Next we show you some examples that constitute noise and they are structures that are extracted by the mapping of these patterns and there are not any terms that characterize that domain.

Pattern Examples of Noise

N estudiante (student), diccionario(dictionary)
NJ diccionario grande (big dictionary)
NPN etapa de trabajo (stage work).

4. Get the Definitory Verbal Patterns

So as to minimize the noise the obtained BP is introduced and a set of DVP to use has been defined.

In (Alarcon et al., 2007), it is shown that the verbs that can operate more as connectors between a term and a definition are conceive and define as well as the prototypical use of the verb to be better than a determiner which is known as ISA relationship.

In a previous study (Alarcon and Sierra 2003) the different definitory verbal patterns were found and they can constitute these verbs, although it is necessary to clarify that, depending on the defined pattern the terms and their definitions can occupy different positions in the constitutive elements.

Based on these two criteria (the verbs are used and the different positions that a term can occupy and its definition in the DVP context) for our methodology we have defined the following DVP:

- BP? DVP + BP?+"como"+*definition*+ BP?
- BP+":"+" DVP "+ *definition*

where:

BP: are obtained in step 3 of the proposed methodology.

DVP: they can be defined taking into consideration the verbs *conceive* and *define*, and the prototypical *is-a* according to the following structures:

SE = Impersonal pronoun *se*
Vaux = Auxiliary verb
VDef_Inf = Definitory verb, impersonal infinitive form.
VDef_Par = Definitory verb, impersonal participle form.
VDef_Con = Definitory verb, personal conjugate form
Pron = pronoun

Definitory verb, impersonal infinitive form.
SE (Pron) VAux VDef_Inf | VAux VDef_Inf (SE | Pron) | VDef_Inf (Pron)
Example: puede definir (se | lo)
Definitory verb, impersonal participle form.
(SE VAux | Vaux{1,2}) Vdef_Par
Example: se ha definido
Definitory verb, personal conjugate form
(SE) VDef_Con
Example: se define

Table 3: Defining Verbal Patterns

In the table above auxiliary verbs (Vaux) can be personal or impersonal forms of any of the above verbs and items in brackets are optional.

With this we are ensuring that the terms that are extracted using these DVP have a defined structure that follows the terms belonging to the domain.

4 Evaluation-Analysis

For the evaluation and analysis processes that follow the proposed methodology the remaining 50% of the selected corpus was used.

Once the corpus is obtained for the evaluation. Table 4 shows some examples of computer science domain terms which are associated to the obtained basic patterns with frequency (F1,F2,...Fn) higher or equal to 80%. (Step 3 of our methodology)

Generally in the BP the precision and recall of the obtained terms from mapping with those patterns were measured. Table 5

Pattern/ Domain	Computer Science
N	computadora (computer) teclado (keyboard)
NJ	programación paralela (parallel programming) sistema operativo (operating system)
NPN	lenguaje de programación (programming language) ingeniería de software(software engineering)

Table 4: List of examples of the basic patterns in the computer science domain

Patterns	Precision (%)	Recall (%)
BP	38,23	97,43

Table 5: Precision and recall values in the BP

We notice that the values of recall for those basic patterns are very good, since most terms have been detected with these structures, however as they are general patterns they introduce much noise, causing the precision values are very low.

In the case of DVP (step 4 for our methodology), we obtain satisfactory precision values, demonstrating that if we include the BP in the DVP we can solve the problem of low precision. However, the covering values are decreasing to a 18%.

Patterns	Precision (%)	Recall (%)
DVP	98,35	18,23

Table 6: Precision and recall values in the DVP

The recall results are low because the definitory verbal patterns only recognize the terms of the corpus that are defined and they do not consider other undefined terms that belong to the domain. Example: A **computer** *is an* equipment which is made up of a CPU and peripherals.

The PVD only extract the term computer and not the terms CPU and peripherals.

4.1 Evaluation of the Obtained Methodology in Other Domains

In order to test the applicability of the proposal methodology in other domains Agricultural Engineering, Veterinary Medicine and Agronomy were selected.

After a validation process we have proved that the terms that characterize these domains correspond to the above basic patterns. Some examples of terminology are shown in Tables 7,8 and 9. Each domain associated respectively with the patterns is also shown.

Pattern/ Domain	Agricultural Engineering
N	agrícola (agricultural)
NJ	maquinaria agrícola (agricultural machinery) producción agropecuaria(agricultural production)
NPN	procesos de poscosecha (post-harvest processes) acidez del suelo (soil acidity) rotación de cultivos (crop rotation)

Table 7: Example list of the obtained basic patterns in Agricultural Engineering domain

Pattern/ Domain	Veterinary Medicine
N	zootecnia (animal husbandry) andrología (andrology)
NJ	medicina veterinaria (veterinary medicine) andrología veterinaria (veterinary andrology)
NPN	transferencia de embriones (embryo transfer)

Table 8: Example list of the obtained basic patterns in Veterinary Medicine domain

Pattern/ Domain	Agronomy
N	Fitotecnia (plant science), hortícola (horticulture)
NJ	producción agrícola (agricultural production) sanidad vegetal (plant health)
NPN	elementos de agroecología (elements of agroecology)

Table 9: List of examples of the obtained basic patterns in Agronomy domain

Similar behavior of the computer science domain corresponded to the results of accuracy and recall in both BP and DVP in each evaluated domain. Table 10 shows the results.

Domain	Patterns	Precision (%)	Recall (%)
Agricultural Engineering	BP	36,34	96,32
	DVP	97,47	20,18
Veterinary Medicine	BP	39,65	98,24
	DVP	98,06	19,56
Agronomy	BP	35,08	96,45
	DVP	96,43	17,18

Table 10: Precision and recall values which were obtained in the domains of Agricultural Engineering, Veterinary Medicine and Agronomy

5 Conclusions and Future Work

In this article we have proposed a methodology for automatic construction of patterns for extracting domain terminology in the Spanish language; it represents a contribution of some importance to this field. The methodology was initially applied to the domain of Computer Science and then was tested in Agricultural Engineering, Veterinary Medicine and Agronomy domains, getting excellent results, showing that it can be applied in any domain of specialty curriculum documents.

In the process of evaluation we have demonstrated that if we only use the BP, we could solve the problem of recall but you know they are very general patterns therefore the problem of accuracy will be affected as well as all nouns, nouns + adjectives, etc will be extracted too.

Incorporating these BP to DVP, we would solve the problem of accuracy, but it is true that the terms of specialty are generally defined in specialized texts, most of these terms are only found in those texts where they are precisely defined, so we could only obtain the terms that are defined in each document.

We propose to use both patterns BP and the DVP due to the fact that the patterns in an extracted system of terminology are an intermediate step in the process, then each extracted system that uses them must validate a set of characteristics either language statistics or semantics that allow them to refine a list of candidates from obtained terms through patterns that were presented here.

As future works we propose to analyze how to combine both sets of patterns (BP and VDP) to obtain the best values of precision and recall. Add new patterns to extract non defined terms in the corpus belonging to the domain and then to use the presented methodology for the creation of an extracted system of terminology that is independent from the domain with the aim of generating a semantic network that can be used in several applications of natural language processing as mentioned above with extracted terms and some linguistic resources EuroWordNet (Vossen, 2001), Babelnet (Navigli, Ponzetto, 2010), DBPedia (S.Auer, 2007) and others.

References

Alarcón, Rodrigo. 2009. Extracción automática de contextos definitorios en corpus especializados. Tesis de Doctorado, Universidad Pompeu Fabra, Barcelona.

Navigli, R. & Ponzetto, S. P. 2010. Babelnet: building a very large multilingual semantic network. In proceedings of the 48th annual meeting of the Association for Computational Linguistics, ACL '10. Stroudsburg, PA, USA, 216–225.

Vossen Piek, 2001. Building a multilingual database with wordnets for several European languages. Language Resources, Language Engineering. 2001.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives, 2007. DBpedia: A Nucleus for a Web of Open Data.

Ellen Riloff. 1993. Automatically constructing a dictionary for information extraction task. In *Procceding of the Eleventh National Conference on Artificial Intelligence.*

Stephend Soderland, David Fisher, Jonathan Aseltine, and Wendy Lehnert. 1995. Crystal: Inducing a conceptual dictionary. In *Proccedings of the Fourteenth International Joint Conference on Artificial Intelligence.*

Ellen Riloff & Jay Shoen. 1995. Automatically acquiring conceptual patterns whitout and annotated corpus. In *Procceding of the Third Workshop on Very Large Corpora*, pages 148-161.

Scott B Huffman. 1995. Learning information extraction patterns from examples. In *IJCAI-95 Workshop on New Approaches to Learning for NLP.*

Neus Catalá & Núria Castell. 1997. Construcción automática de diccionarios de patrones de extracción de información.

Dubuc, R. & Lauriston, A. 1997. "Terms and Contexts". In *Handbook of Terminology Management*. Volume 1. Wright, S.E. and Budin, G. (eds). Amsterdam/Philadelphia: John Benjamins Publishing Company, 80-87.

De Bessé, Bruno. 1991. Le Contexte Terminographique. *Meta* 26(1):111-120.

Auger, Alain. 1997. Repérage des énoncé d'intérêt définitoire dans les bases de données textuelles. Tesis de doctorado, Neuchâtel, Universidad de Neuchâtel.

Pearson, Jennifer. 1998. *Terms in Context*, Philadelphia, John Benjamins.

Meyer, Ingrid. 2001. Extracting a knowledgerich contexts for terminography: A conceptual and methodological framework. In *Recent Advances in Computational Terminology*, edited by Bourigault, D.; Jaquemin, C. & L'Homme, M.C. Philadelphia: John Benjamins.

Alarcón, R. & Sierra, G. 2003. El rol de las predicaciones verbales en la extracción automática de conceptos, Estudios de Lingüística Aplicada 38, México DF, Universidad Nacional Autónoma de México-Centro de Enseñanza en Lenguas Extranjeras, pp. 129-144.

Sierra, Gerardo. 2009. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. Universidad Nacional Autónoma de México.

Hassan Saneifar, Stephane Bonniol, Anne Laurent, Pascal Poncelet and Mathieu Roche, 2009. Terminology Extraction from Log Files.

Gerardo Sierra, Alfonso Medina, Rodrigo Alarcón, César A. Aguilar, 2006. Towards the Extraction of Conceptual Information from Corpora.