Syntactic-Semantic Extraction of Patterns Applied to the US and European Patents Domain

Anabel Fraga, Juan Llorens, Eugenio Parra, Leticia Arroyo and Valentín Moreno
Computer Science Department, Carlos III of Madrid University, Av. Universidad 30, Leganés, Madrid, Spain

Keywords: Indexing, Ontologies, Knowledge, Patterns, Reuse, Retrieval, Patents, US Patents, European Patents.

Abstract: Nowadays, there are many scientific inventions referring to any topic like medicine, technology, economics, finance, banking, computer science, and so on. These inventions are suggested as patents to the agencies working in US and Europe for the registration and revision of the patent applications. But, the job of reviewing the patents might be complicated because every day the quantity of it is bigger and bigger. And also, the amount of work dedicated writing a proper application might be intricate and needs several revisions from investor and examiners. This revision job might have costs for the inventor because they don't know the proper language for writing the application in the formal mode used. As part of a solution, one approach to minimize the impact of this fact and increase the success of the reviewing process is aid the human reviewer and also inventors with a set of patterns created using Natural Language Processing techniques that accelerate the review just looking in the massive set of registration any similar one already patented and on the other hand aid the inventor writing in the formal manner the application.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

The process for applying to an Intellectual Property protection, as patents, might be complex and reviewing your invention is really patentable must be approved and check by an examiner. Also, the language used to specify the invention in the application is specific to this domain.

If it could be possible to extract a set of patterns aiding the inventor and examiner in the process of construction of the application and also reviewing if the inventions could be already patented, the process of patenting could be improved in two different viewpoints.

Christopher Manning states in his book that: "People write and say lots of different things, but the way people say things - even in drunken casual conversation - has some structure and regularity."(Manning, 1999)

The important aspect in here is to ask ourselves: how do people write? Nowadays, researchers conduct investigations using natural language processing tools, generating indexing and semantic patterns that help to understand the structure and relation of how writers communicate through their papers.

This project will use a natural language processing system which will analyze a corpus of patents acquired from the open repository of the US patent and European patent Agencies. The documents will be processed by the system and will generate simple and composed patterns. These patterns will give us different results which we can analyze and conclude the common aspects the documents have even though they are created by different authors but are related to the same topic. (Alonso et al., 2005) The study uses as center of the study an ontology created in a national founded project for Oncology and it has been extended with general terms of public health.

The remainder of this paper is as follows: section 2 includes the state of the art and related work of the main topics of research, section 3 includes the summary of the methodology; section 4 summarizes the results, and finally conclusions.

2 STATE OF THE ART AND RELATED WORK

2.1 Information Reuse

Reuse in software engineering is present throughout the project life cycle, from the conceptual level to the definition and coding requirements. This concept is feasible to improve the quality and optimization of

Fraga, A., Llorens, J., Parra, E., Arroyo, L. and Moreno, V.
Syntactic-Semantic Extraction of Patterns Applied to the US and European Patents Domain.
DOI: 10.5220/0006098600360043
In *Proceedings of the 7th International Workshop on Software Knowledge (SKY 2016)*, pages 36-43
ISBN: 978-989-758-202-8
Copyright © 2016 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

the project development, but it has difficulties in standardization of components and combination of features. Also, the software engineering discipline is constantly changing and updating, which quickly turns obsolete the reusable components (Llorens, 1996).

At the stage of system requirements reuse is implemented in templates to manage knowledge in a higher level of abstraction, providing advantages over lower levels and improving the quality of the project development. The patterns are fundamental reuse components that identify common characteristics between elements of a domain and can be incorporated into models or defined structures that can represent the knowledge in a better way.

2.2 Natural Language Processing

The need for implementing Natural Language Processing techniques arises in the field of the human-machine interaction through many cases such as text mining, information extraction, language recognition, language translation, and text generation, fields that requires a lexical, syntactic and semantic analysis to be recognized by a computer (Cowie et al., 2000). The natural language processing consists of several stages which take into account the different techniques of analysis and classification supported by the current computer systems (Dale, 2000).

1) Tokenization: The tokenization corresponds to a previous step on the analysis of the natural language processing, and its objective is to demarcate words by their sequences of characters grouped by their dependencies, using separators such as spaces and punctuation (Moreno, 2009). Tokens are items that are standardized to improve their analysis and to simplify ambiguities in vocabulary and verbal tenses.
2) Lexical Analysis: Lexical analysis aims to obtain standard tags for each word or token through a study that identifies the turning of vocabulary, such as gender, number and verbal irregularities of the candidate words. An efficient way to perform this analysis is by using a finite automaton that takes a repository of terms, relationships and equivalences between terms to make a conversion of a token to a standard format (Hopcroft et al., 1979). There are several additional approaches that use decision trees and unification of the databases for the lexical analysis but this not covered for this project implementation (Trivino et al., 2000).
3) Syntactic Analysis: The goal of syntactic analysis is to explain the syntactic relations of texts to help a subsequent semantic interpretation (Martí et al., 2002), and thus using the relationships between terms in a proper context for an adequate normalization and standardization of terms. To incorporate lexical and syntactic analysis, in this project were used deductive techniques of standardization of terms that convert texts from a context defined by sentences through a special function or finite automata.
4) Grammatical Tagging: Tagging is the process of assigning grammatical categories to terms of a text or corpus. Tags are defined into a dictionary of standard terms linked to grammatical categories (nouns, verbs, adverb, etc.), so it is important to normalize the terms before the tagging to avoid the use of non-standard terms. The most common issues of this process are about systems' poor performance (based on large corpus size), the identification of unknown terms for the dictionary, and ambiguities of words (same syntax but different meaning) (Weischedel et al., 2006). Grammatical tagging is a key factor in the identification and generation of semantic index patterns, in where the patterns consist of categories not the terms themselves. The accuracy of this technique through the texts depends on the completeness and richness of the dictionary of grammatical tags.
5) Semantic and Pragmatic Analysis: Semantic analysis aims to interpret the meaning of expressions, after on the results of the lexical and syntactic analysis. This analysis not only considers the semantics of the analyzed term, but also considers the semantics of the contiguous terms within the same context. Automatic generation of index patterns at this stage and for this project does not consider the pragmatic analysis.

2.3 RSHP Model

RSHP is a model of information representation based on relationships that handles all types of artifacts (models, texts, codes, databases, etc.) using a same scheme. This model is used to store and link generated pattern lists to subsequently analyze them using specialized tools for knowledge representation (Llorens et al., 2004). Within the Knowledge Reuse

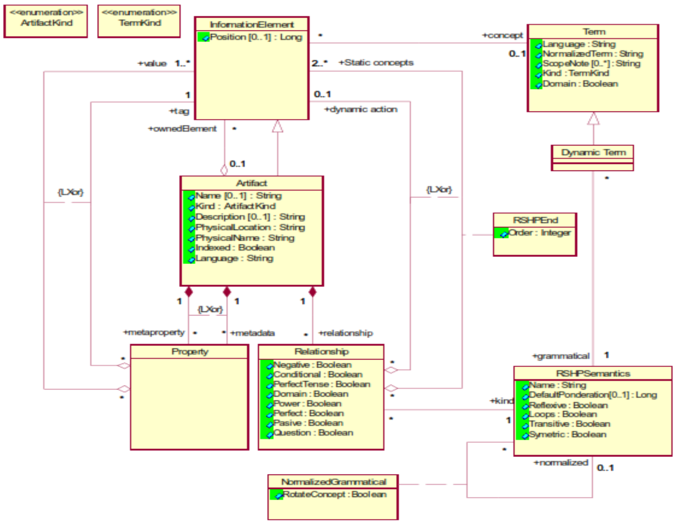


Figure 1: RSHP information representation model. (Alonso et al., 2005).

Group at the University Carlos III of Madrid RSHP model is used for projects relevant to natural language processing. (Gomez-Perez et al., 2004) (Thomason, 2012) (Amsler, 1981) (Fraga, 2010) (Suarez et al., 2013). The information model is presented in Figure 1.

3 METHODOLOGY

The objective of this research is to perform the extraction of syntactic-semantic patterns found within documents on patents.

Patent documents are written by experts, therefore we are saying that we will have very well written documents and high quality grammatical.

When the investigation is complete, we have a list sorted by frequency patterns (See Figure 2). We will know the syntactic-semantic patterns that are most used when writing a patent.

In addition to patterns, the most recurrent words are known, we will identify the most common words in the patterns documents.

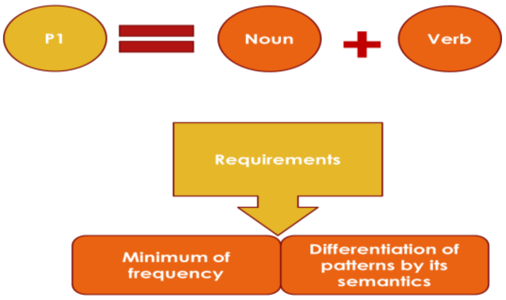


Figure 2: Frequency of patterns.

The phases defined here are needed to meet the objectives:

PHASE 1: Search for patent sources where they can download patents documents public and registered in PDF format. The documents must be converted to TXT format using pdf2txt. Pdf2txt is a program available in internet.

PHASE 2: Download at least about 500 documents.

PHASE 3: Convert the PDF documents to TXT using the pdf2txt program.

PHASE 4: Get WordNet dictionary to form the ontology. This phase can be performed in parallel to steps 1, 2 and 3.

PHASE 5: Manage the ontology with a software for managing ontologies in the industrial domain, *KnowledgeMANAGER*. Adding vocabulary obtained in phase 4.

PHASE 6: Add the new ontology in *BoilerPlates* tool[1], a tool for detecting patterns (Boilerplates in its most initial form) in a set of documents.

PHASE 7: Define study scenarios and using ontology created, generating patterns with the *BoilerPlates* tool.

PHASE 8: TXT documents will be included one by one on the *BoilerPlates* tool, with this first step in the tool will generate the basic patterns.

[1] *BoilerPlates* Tool is a software developed in a PhD Dissertation (Parra, 2016) in order to generate patterns of text using Natural Language Processing solutions.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

PHASE 9: Representing one to one each scenario in *BoilerPlates* tool and start pattern generation.

PHASE 10: Analyze the results obtained by scenario.

PHASE 11: Analyze and compare the results of all scenarios.

In this work a syntactic-semantic analysis is performed, of a sample of registered patents and made public, through an ontology based on natural language words.
To get a larger sample of patent documents to analyze them, it has decided to use English as the language of analysis. Therefore all patents that are used in this investigation will be written in the English language.
All patents are search in Internet and document must be PDF formats.
It does not establish any particular subject, and not any particular area of investigation, the investigation developed here is valid for all subjects.
We have two samples of patents, on one hand analyze documents of the United States Patent and Trademark Office, we have 359 documents, and secondly analyze documents of the European Patent Office, we have 379 documents Europeans different.
The study will be made with over 700 patent documents, all documents be analyzed with the *BoilerPlates* tool.
The ontology that includes the boilerplates tool, will be managed with the *KnowledgeMANAGER* tool of REUSE Company. The vocabulary will form the ontology is providing by WordNet.
WordNet is used as a basis for the ontology of data recovery, we will have a language general controlled (not specialized by subject) and to language English. Into the WordNet we obtain nouns, verbs, adjectives and adverbs.
The investigation done here is interesting because we discover how the pattern of professional experts document their investigations, findings and studies.
Here art to documentation is analyzed, so important it is to have an idea as important is knowing it registered.
The patterns that are obtained in the investigation may be useful in the future to guide the new professionals in the time of writing or searching for similar patents.

4 RESULTS

The Scenarios followed in the experiments are:

Scenario 1:

- Sample USPTO (United States Patent and Trademark Office) patents.
- All grammatical categories available are used
- Use a minimum frequency of 1 to create patterns
- Differentiate patterns by their semantics is disabled

Scenario 2:

- Sample USPTO patents.
- Use all grammatical categories.
- Use a minimum frequency of 1 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 3:

- Sample USPTO patents.
- Use all grammatical categories.
- Use a minimum frequency of 20 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 4:

- Sample EPO (European Patent Office) patents.
- Use all grammatical categories.
- Use a minimum frequency of 20 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 5:

- Sample USPTO patents
- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 6:

- Sample EPO patents.
- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is enabled.

Scenario 7:

- Sample USPTO patents.
- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is disabled.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Scenario 8:

- Sample EPO patents
- Use all grammatical categories.
- Use a minimum frequency of 100 to create patterns.
- Differentiate patterns by their semantics is disabled.

In addition to analyzing each scenario separately, comparisons between 1-2, 3-4, 5-6 and 7-8 pairs were carried out to compare the two sources of information used.
Comparative analysis - all in common scenarios will also be made to draw general conclusions to all the analyzed scenarios.

Basic patterns:

After the basic patterns were created, all the sentences from the text documents were analyzed and to each of the words (known in the database as token text) a term tag or syntactic tag was assigned with the help of the tables Rules Families and Vocabulary in the Requirements Classification database.
You may find the most repeated words in the domain of documents in the Basic patterns table. The most repeated words in grammatical categories such as nouns, verbs and nouns coming from the ontology we used.

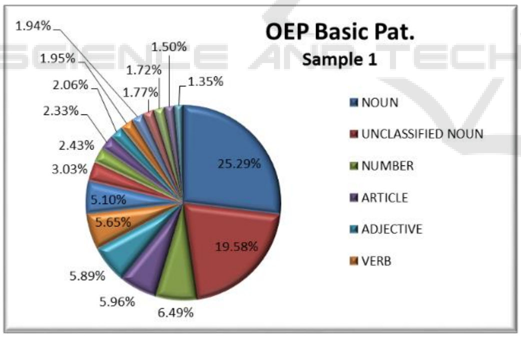


Figure 3: Basic pattern results. OEP Sample 1.

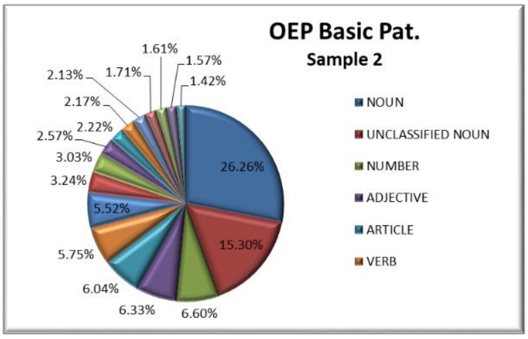


Figure 4: Basic pattern results. OEP Sample 2.

It can be seen that in both cases there is little difference between the two samples , more number of words in sample 1 than in sample 2, but not shown in the percentage of appearance in each of the grammatical categories.

USPTO vs OEP:

Comparing the two results, we see that the number of grammatical categories exceeding 1 % is the same in both, but with slight differences. In USPTO items are the third most repeated grammatical category, while EPO are the numbers in this position. In the latter, the repetition of basic patterns is not rated much higher.
We see in the figure below the comparative representation of the 17 most repeated grammatical categories.

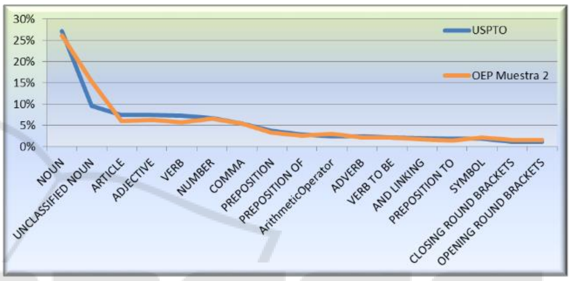
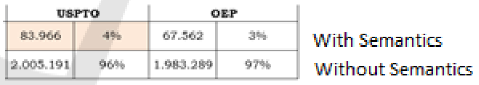


Figure 5: USPTO vs OEP.

Semantics:

Semantics is present within the basic patterns but in a very limited way. We met a little more than semantics within American samples.



USPTO		OEP		
83.966	4%	67.562	3%	With Semantics
2.005.191	96%	1.983.289	97%	Without Semantics

In the next chart you can see the semantics that more appears in both samples:

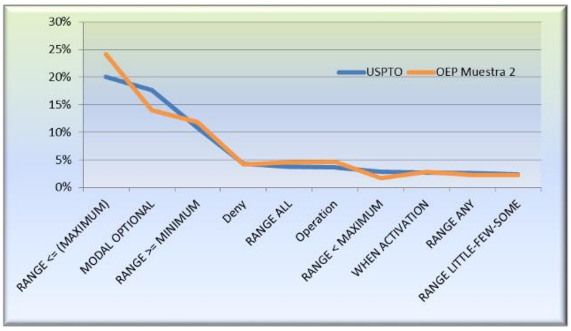


Figure 6: Basic patterns: Semantics.

The stage 1, 2, 3, 5 and 7 are all made with American patent documents.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

For the rest, we can conclude for the US shows the following is true:

- A higher minimum frequency , fewer patterns
- A higher minimum frequency, the lower the semantic obtained.
- Differentiate by semantic patterns is a better practice to know the real semantics being used when writing sentences. Otherwise, for the same pattern, which can adopt semantics could be anyone.
- Not differentiate by semantic results in increased number of patterns, but with fewer sub-patterns that form.

Scenarios 4, 6 and 8 are carried out with sample documents of European patents.

We can conclude, for the European shows the following is true:

- A higher minimum frequency , fewer patterns
- A higher minimum frequency, the lower the semantic obtained.
- Differentiate by semantic patterns is a better practice to know the real semantics being used when writing sentences. Otherwise, for the same pattern, which can adopt semantics could be anyone.
- Not differ semantics resulting greater number of patterns, and the number of sub patterns is very similar.

After the analysis of the US patents documents and European patents documents we can conclude the following:

The basic patterns obtained are independent of the frequency and the selection of grammatical categories in the boilerplates tool. All basic patterns are common within the same sample.

In the boilerplates tool, the higher the minimum frequency used, is less the number of patterns obtained and is shorter the time necessary to obtain them.

Differentiation has been made by their semantic patterns in the minimum frequencies of 1, 20 and 100 to US samples, and 20 and 100 for European samples. For frequency 1 it has not been possible to obtain results due to the high volume of information that we have handled. More than 25 days after running the tool, it has had to reject frequency 1 for the study. About the other two frequencies, we can say that the higher the frequency the number of patterns obtained is less.

Patterns are calculated without differentiation of semantics for the minimum frequencies of 1 and 100 with US sample. It is also calculated with the European sample for the minimum frequency of 100, without differentiation patterns by their semantics. It can be concluded that the same patterns are obtained with different semantics.

With increasing frequency we lose patterns that have longer decomposition. Because the number of repetitions is less.

After using different frequencies to generate patterns in boilerplates, we can say that the intermediate frequency is what has given us the best results.

In both samples the unclassified words are very present.

The patterns obtained in all scenarios can assist the writing for any user who need to write a patent.

After the investigation, with the knowledge obtained now, we can give some recommendations to people who will do a similar study in the future.

The ontology can be improved, the ontology has 73 grammatical categories to define their vocabulary. For this project has not been completed because all the most important words are covered. The pending grammar to define are the type of punctuation, dates, email, arithmetic symbols, acronyms, etc. The undefined categories are shown in Table 8.

For future projects, scenarios of using a minimum frequency of 100 can be applied to search which is the minimum frequency that will create zero patterns.

It is possible create a new analysis with minimum frequency greater than 100, because we obtained patterns where their repetition frequency is greater than 100. But before begin studies with a higher minimum frequency, we recommend you should not consider words that do not correspond to a grammar of the ontology.

After ending all scenarios and analyzing results, we can conclude that authors writing papers about a same topic (in this case, genetic engineering) have similarity in how they write. They use a similar vocabulary and appropriate terms which makes the reading easier. Some additional enterprises need observation and intelligence.

5 CONCLUSION

After the analysis of documents of US patents and European patents we can conclude the following:

- The basic patterns obtained are independent of the frequency and the selection of grammatical categories in the boilerplates tool. All basic patterns are common within the same sample.

- The higher the frequency used in the boilerplates tool, the smaller the number of patterns obtained and less time needed to obtain them.
- Differentiation has been made by their semantic patterns in the minimum frequencies of 1, 20 and 100 for American samples, and 20 and 100 for European samples. To frequency 1 it has not been possible to obtain results due to the large volume of information we have handled. After more than 25 days running the tool, it has had to dismiss frequency 1 for the study. On the other two frequencies, we can say that the higher the frequency the number of patterns obtained is lower.
- Patterns are calculated without differentiation of semantics for the minimum frequencies of 1 to 100 with American shows. It is also estimated with the European sample for the minimum rate of 100, without differentiation patterns by their semantics. It can be concluded that the same patterns are obtained with different semantics.
- By increasing the frequency lose patterns that have greater depth of decomposition. Since your number of repetitions is less.
- After using different frequencies to generate patterns in boilerplates, we can say that the intermediate frequency is what has given us better results.
- In both samples not rated names is very present.
- The patterns obtained in all scenarios may be of assistance to those who need to write a patent.

After the investigation, with the knowledge now acquired, we can give some recommendations who faces a future in a similar study.

- The ontology can be improved, it has 73 labels for outstanding grammatical categories to define their vocabulary. For this project has not been completed because all the most important words are covered. The slopes are grammars to define the type of punctuation, dates, email, arithmetic symbols, acronyms, etc. These categories may be undefined in Table 8.
- There have been many token which are classified under the label "UNCLASSIFIED NOUN". For these cases we see three action plans:
    - Or they could analyze them and give them all a grammatical category if possible, so finding patterns would be more accurate.
    - If it is not possible to assign a particular category, you have to look at the possibility of eliminating all words and symbols are not classifiable.
    - When generating patterns with boilerplates tool not consider the label "UNCLASSIFIED
- The documents have been used in this analysis can be improved by converting PDF to TXT performed in this process has been lost information. Documents with images are those that have lost more information.

It is possible to perform the analysis when frequency is greater than 100, since we obtained patterns where the repetition frequency is greater than 100. But before studies with higher minimum frequency, it is recommended not to consider if the terms do not correspond to a grammar of the ontology.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the european union’s seventh framework program (fp7/2007-2013) for crystal – critical system engineering acceleration joint undertaking under grant agreement n° 332830 and from specific national programs and / or funding authorities.

Also, the research leading to this paper has received funding from the AMASS project (H2020-ECSEL grant agreement no 692474; Spain’s MINECO ref. PCIN-2015-262).

REFERENCES

Abney, Steven. Part-of-Speech Tagging and Partial Parsing, S. young and G. Bloothooft (eds.) Corpus-Based Methods in Language and Speech Processing. An ELSNET book. Bluwey Academic Publishers, Dordrecht. 1997.

Alonso, Laura. Herramientas Libres para Procesamiento del Lenguaje Natural. Facultad de Matemática, Astronomía y Física. UNC, Córdoba, Argentina. 5tas Jornadas Regionales de Software Libre. 20 de noviembre de 2005. Available in: http://www.cs.famaf.unc.edu.ar/~laura/freeNLP

Amsler, R. A. A taxonomy for English nouns and verbs. Proceedings of the 19th annual Meeting of the Association for Computational Linguistic. Stanford,

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

California, 1981. Pp. 133-138.

Carreras, Xavier. Márquez, Luis. Phrase recognition by filtering and ranking with perceptrons. En Proceedings of the 4th RANLP Conference, Borovets, Bulgaria, September 2003.

Cowie, Jim. Wilks, Yorick. Information Extraction. En DALE, R. (ed). Handbook of Natural Language Processing. New York: Marcel Dekker, 2000. Pp.241-260.

Dale, R. Symbolic Approaches to Natural Language Processing. En DALE, R (ed). Handbook of Natural Language Processing. New York: Marcel Dekker, 2000.

Fraga, Anabel. A methodology for reusing any kind of knowledge: Universal Knowledge Reuse. PhD Disertation. Universidad Carlos III de Madrid, 2010.

Gómez-Pérez, Asunción. Fernando-López, Mariano. Corcho, Oscar. Ontological Engineering. London: Springer, 2004.

Hopcroft, J. E. Ullman, J. D. Introduction to automata theory, languages and computations. Addison-Wesley, Reading, MA, United States. 1979.

Llorens, J., Morato, J., Genova, G. Rshp: An Information Representation Model Based on Relationships. In Ernesto Damiani, Lakhmi C. Jain, Mauro Madravio (Eds.), Soft Computing in Software Engineering (Studies in Fuzziness and Soft Computing Series, Vol. 159), Springer 2004, pp. 221-253.

Llorens, Juan. Definición de una Metodología y una Estructura de Repositorio orientadas a la Reutilización: el Tesauro de Software. Universidad Carlos III. 1996.

Manning Christopher, "Foundations of Statistic Natural Language Processing", Cambridge University, England, 1999, 81

Martí, M. A. Llisterri, J. Tratamiento Del Lenguaje natural. Barcelona: Universitat de Barcelona, 2002. p. 207.

MORENO, Valentín. Representación del Conocimiento de proyectos de software mediante técnicas automatizadas. Anteproyecto de Tesis Doctoral. Universidad Carlos III de Madrid. Marzo 2009.

Parra, Eugenio. Metodología orientada a la optimización automática de la calidad de los requisitos. PhD Disertation. Universidad Carlos III de Madrid, 2016.

Poesio, M. Semantic Analysis. En DALE, R. (ed). Handbook of Natural Language Processing. New York: Marcel Dekker, 2000.

Rehberg, C. P. Automatic Pattern Generation in Natural Language Processing. United States Patent. US 8,180,629 B2. May 15, 2012. January, 2010.

Riley, M. D. Some applications of tree-based modeling to speech and language indexing. Proceedings of the DARPA Speech and Natural Language Workshop. California: Morgan Kaufmann, 1989. Pp. 339-352.

Suarez, P., Moreno, V., Fraga, A., Llorens, J. Automatic Generation of Semantic Patterns using Techniques of Natural Language Processing. SKY 2013: 34-44.

Thomason, Richmond H. What is Semantics? Version 2. March 27, 2012. Available in: http://web.eecs.umich.edu/~rthomaso/documents/general/what-is-semantics.html

Triviño, J.L. Morales Bueno, R. A Spanish POS tagger with variable memory. In Proceedings of the Sixth International Workshop on Parsing Technologies (IWPT-2000). ACL/SIGPARSE, Trento, Italia, 2000. pp. 254-265.

Weischedel, R. Metter, M. Schwartz, R. Ramshaw, L. Palmucci, J. Coping with ambiguity and unknown through probabilistic models. Computational Linguistics, vol. 19, pp. 359-382.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS