EXTRACTION OF PATTERNS USING NLP: GENETIC DEAFNESS[1]

Anabel Fraga[1], Javier Garcia[1], Eugenio Parra[1], Valentín Moreno[1]

[1] *Computer Science Department, Carlos III of Madrid University*
Av. Universidad 30, Leganés, Madrid, Spain
{afraga, eugenio.parra@kr.inf.uc3m.es, vmoreno}@inf.uc3m.es

Keywords: indexing, ontologies, knowledge, genetic deafness, medicine, reuse, retrieval, conceptual text integrity.

Abstract: In the domain of Genetic Deafness in medicine, it is important to detect some patterns. Medical doctors needs to search and deal with information from diverse sources and it is important to be able to cross information between sources. As part of a solution, one approach to minimize the impact of this lack and increase the success of the retrieval process crossing diverse sources of information laid in the use of Natural Language Processing techniques permitting conceptual integrity of text.

1 INTRODUCTION

In this research project, we solved the following problem: is it possible to make a computer recognize linguistic patterns in a document? For example, when reading the sentence:
“The patient had a temperature of 102F at Saturday”
With syntactic tokens: “Determiner + Noun + Verb + Preposition + Noun + Preposition + Noun + Preposition + Noun”
Is it possible for the computer to understand it both syntactically and semantically? Moreover, once this information is acquired, is it possible to generate patterns that allow recognizing similar sentences throughout a document?

With this problem as a starting point, this project is related to the studies of Natural Language Processing (NLP). For that, by using vocabulary from a specific domain, the computer will read a number of documents and automatically generate all linguistic patterns contained in them. Therefore, there are two main objectives:

- Acquire the terminology from a specific domain.

[1] DOI: 10.18293/SEKE2017-204

- Perform the linguistic pattern learning process by reading and analyzing a number of documents from the same domain.

The chosen domain for this project was “Genetic Deafness” (medicine), and the language used was English.

This paper is a summary of the project, which covers the followed work procedure and the results of the experimentation, as well as the conclusions reached after the process was completed.

There were two potential motivations for this research project:

- Effective information extraction from a text document. When a human reads a sentence, he or she extracts a piece of the information that he or she considers useful. See the previous example: “The patient had a temperature of 102F at Saturday”. From this sentence, we can retain the information related to the temperature (102F) and/or the time when the measure was taken (at Saturday).

For a computer, this “understanding” process is difficult. There are no trivial algorithms that, after reading the sentence, can storage the knowledge received about the patient’s temperature and the day the measure was taken. A recurrent solution for this problem is to use templates and file formats, which allows the computer to receive all the information without actually understanding it. This, however, forces the information to be pre-processed so it can meet software specifications, which consumes more time and effort.

A previous learning of linguistic patterns may solve this problem. If the computer is able to recognize the example sentence above as a learned pattern, then it would realize that the seventh word indicates the temperature (102F) and the last one indicates the day (Saturday).

Therefore, the pattern analysis could be useful to acquire all relevant information from a document of variable length.

- Writing style guide development. In several fields, like law or medicine, there are specific writing styles to follow. Therefore, the results from this project in a specific domain can be useful in a didactic way. For example: which word families are more frequent in the domain? Which linguistic sequences are more common? Is there a recognizable conceptual integrity of the text in a higher-level than individual concepts and even recurring patterns? And so on.

By studying the linguistic patterns from a domain, we can contribute to the learning of the writing style expected in a document, in order to show all the information in an organized state. This is also helpful for computers: when the writing style is standardized, the pattern recognition is easier and, therefore, the information retrieval is easier and more efficient.

2 STATE OF THE ART AND RELATED WORK

2.1 How to Define Conceptual Integrity of Texts?

Conceptual Integrity has been defined by Brooks [ref], mainly in the context of software systems design and development. In this paper we deal with texts, which might be software assets, such as software documentation in several domains, but are certainly different than software system design tools or runnable software itself.

Since our main research goal of this paper is to enable mining of relevant texts for a given domain – in our case study it is “Genetic Deafness” – one asks: what is the relevant definition of “Conceptual Integrity” in this context?

A preliminary proposal for such a definition would include the following steps:

- Choose a basic set of concepts to characterize the desired texts in the desired domain;
- Select a set of patterns – to be obtained and explained later on in the research of this paper – containing the basic set of concepts;
- Determine reasonable minimal thresholds, to check whether the appearance of the chosen concepts and patterns are above these reasonable thresholds.

In our investigation, as future work, we shall make extensive tests of these steps for certain chosen domains.

2.2 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 needed to improve the quality and optimization of 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).

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.

2.3 Natural Language Processing

The need for implementing Natural Language Processing techniques (see Figure 1) 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).

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 (Hopcroft et al., 1979).

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.

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 (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. One of the main challenges NLP has tried to overcome since its beginnings is the problem of ambiguity.

3 METHODOLOGY

The methodology of this project is based on the usage of concepts related to genetic deafness in order to recognize the most frequent patterns from a number of documents while facing the problems and limitations from the Natural Language Processing.

In this project, the BoilerPlates tool has been used. This tool, developed by the Carlos III de Madrid research team (Knowledge Reuse Research Team), can perform the analysis of linguistic patterns given a previously learned set of concepts and grammar categories. Its database already contains vocabulary that is not related to any domain (such as the verb "to be", prepositions, adverbs…). However, to effectively use this tool in a certain domain, the database must be updated with domain-specific concepts prior to the pattern extraction.

The methodology used for this project is divided in the following steps:

1. Acquisition and processing of domain documents.
2. Extraction of related terminology.
3. Adaptation of terminology to BoilerPlates.
4. Using the BoilerPlates tool. It has been used. This tool, developed by the Carlos III de Madrid research team (Knowledge Reuse Research Team), can perform the analysis of linguistic patterns given a previously learned set of concepts and grammar categories. Its database already contains vocabulary that is not related to any domain (such as the verb "to be", prepositions, adverbs…). However, to effectively use this tool in a certain domain, the database must be updated with domain-specific concepts prior to the pattern extraction.
5. Extraction of results.
6. Exposition and analysis of results.

4 RESULTS

Eight scenarios were identified for this study, with minimal frequencies from 1 to 20without or with semantic distinction. Once the scenarios were individually analysed, the combined results were as it follows:

Number of patterns identified

As expected, the lower the minimal frequency (MF) value chosen, the more patterns are considered in the identification process.

The growth of the number of patterns found is inversely exponential: in the experiments with MF established to 1, about 140.000 patterns were found.

When raising the MF to 5, the number drops at 14.500 on average, which is about a 90% decrease on patterns found. On the experiments with MF set to 10 and 20, the number of patterns were approximately 5.600 and 2.800 respectively, which is a considerable drop but less severe than before.

Patterns related to genetic deafness

After analysing the results as a whole set of patterns, a subset was created for each scenario, containing only those patterns which has at least one concept related to genetic deafness ("domain patterns").

This subset was a minority in every studied case. For all scenarios, about 10 to 16% of the patterns contained at least one concept from the domain. This was predictable, as most of the words contained in one document about a certain topic are not terminology from the domain, such as prepositions, numbers or unrelated nouns. It is important for conceptual integrity.

Conclusions

As Manning said and we are investigating in our hypothesis, it is possible to get structure and regularity in a set of documents in order to achieve a regular pattern to write or suggest better manners to prepare documents with a higher complexity. Natural language is complex and its regularity depends on the domain we are dealing with. The more regular and mature the domain, the best it is to extract a set of patterns. The more structure the more effective the set of patterns.

Once the analysis of the results was completed, we reached some of the following conclusions regarding the pattern identification and studio in the domain of genetic deafness:

1. The minimal frequency value made a considerable impact in the analysis. The higher the value, the lower the number of identified patterns, since a sequence must reach a bigger number of repetitions until it is considered a pattern.
2. On the matter of semantic distinction: The first experiments showed that there was little difference whether applying semantics or not. However, when the weighted study was completed, it was discovered that semantics played a considerable role during the analysis, such as which patterns were considered the most valuable.
3. Of all patterns discovered, around 10% to 16% had at least one domain concept.
5. The most frequent concepts that are related to "genetics" were those that are about gene manipulation and mutation.

References

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.

FRAGA et al. SYNTACTIC-SEMANTIC EXTRACTION OF PATTERNS APPLIED TO THE US AND EUROPEAN PATENTS DOMAIN. SKY2016 / IC3K2016. Portugal, 2016.

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.

MARTIN, James N. 1996. Systems Engineering Guidebook: A Process for Developing Systems and Products. CRC Press, Inc.: Boca Raton.

MORENO, Valentín, Pablo Miguel Suárez, Anabel Fraga, Juan Llorens, and Eugenio Parra. 2013. Método de generación de patrones semánticos. PCT/ES2013/070638, issued 2013.

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. 2016. Metodología orientada a la optimización automática de la calidad de los requisitos. PhD

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

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.