Modeling and Implementation of a Spanish Question Answering System in Python

Jesus Iram Castañeda-Rivera[1], Roboam Casanova-Castro[2] and Rogelio Dávila-Pérez[3]

[1] Graduate student
Master in Computer Science
Universidad Autónoma de Guadalajara
blinketito@hotmail.com

[2] Graduate student
Master in Computer Science
Universidad Autónoma de Guadalajara
zr.rollo@gmail.com

[3] Researcher-professor
Graduate Division in Computer Science
Universidad Autónoma de Guadalajara
rdav90@gmail.com

Abstract

Due to the huge amount of information circulating on the Internet in recent years, identifying relevant information in real time is becoming increasingly difficult. We have different ways to search for information in different sources e.g. in Internet we have the search engines, such as Google or Facebook. With database systems we use the SQL language. Natural language processing has emerged as a possibility to recovering information from unstructured sources in Internet and Question Answering Systems based in natural languages such as English or Spanish are interesting alternatives to the SQL language, they are easy to use and allow for information retrieval without the need to learn an artificial language such as SQL. The implementation of natural language question answering systems is one of the most complex and interesting challenges in Natural Language Processing. The present document propose a model for interpreting Spanish questions to retrieve information from a relational database, the system is based on NatLIn, a Spanish natural language front-end developed in Prolog for answering questions from a geographical database. The system, that we identify as PyNatLIn was developed in Python, a high-level programming language that posses convenient features that allowed for natural language syntactic and semantic analysis and the possibility of interconnection with commercial relational database systems. The purpose of the paper is to present the design and implementation of PyNatLIn system and the adequacy of Python for natural language processing.

Keywords: Natural Language Processing, Question Answering Systems, and Information Retrieval.

1. Introduction

Natural Language Processing (NLP) is a field which relates artificial intelligence, logic and linguistics which is concerned with the making the computers to process human

languages, that is, the analysis, interpretation, understanding and generation of languages that humans use. The development of natural language question answering systems for knowledge bases is one of the most useful and interesting applications of NLP. Several systems have been developed such as the one developed by Veronica Dahl [9] and one the best, Pereira's Chat-80 [7], which opened the possibility of implementing a semantically consistent formal interpretation of English. Unfortunately few attempts have been done for building Spanish natural language front-ends for relational databases. The system presented in this document, PyNatLIn, is based in NatLIn [1, 3, 5] a Spanish question answering system built in Prolog. NatLIn resemble Chat-80's performance but instead of using English as a central language is oriented toward the analysis and processing of Spanish questions. PyNatLIn, makes explicit and modifies the Spanish grammar used by NatLIn creating a framework for language analysis and interpretation using predicate logic as a semantic representation, adding NLP procedures for allowing the user to retrieve information stored in a database by typing requests articulated in Spanish. The following examples show the communication between the user and the NatLIn original system:

PREGUNTA: ¿Qué países colindan con México?
(QUESTION: What countries border with Mexico?)

[belice][guatemala][estados_unidos]

PREGUNTA: Cuáles son las capitales de los países que colindan con el mar Rojo?
(QUESTION: What are the capitals of the countries that border the Red Sea?)

[jerusalem][amman][riad][sana][el_cairo][addis_ababa][jartum][asmara][yibuti]

If we want to know not only the capital city but the actual country as well we modify the question by adding and additional interrogative pronoun:

PREGUNTA : Cuáles son las capitales de que países que colindan con el mar Rojo?
(QUESTION: What are the capitals of which countries that border the Red sea?)

[jerusalem,israel][amman,jordania][riad,arabia_saudita][sana,yemen][el_cairo,egipto]
[jartum,sudan][asmara,eritrea][yibuti,yibuti]

In section 2, we propose introduction and justification for the using of Python as a framework for development; in section 3, we introduce a Natural Language framework for the analysis of Spanish; in section 4, the implementation of PyNatLIn is proposed and then some conclusions are presented in section 5.

2. Python Framework

The Python scripting language is an interpreted general-purpose high-level programming language. It also contains abundant and versatile libraries, among them, an important library for natural language processing called NLTK, it is a leading platform for building Python programs to work with natural language data. It provides basic classes for representing data relevant to natural language processing; standard interfaces for performing tasks such as part-of-speech tagging, syntactic parsing, and text classification; and standard implementations for each task which can be combined to solve complex problems.[4]. Another important feature of Python is that it allows the possibility of testing the system with

commercial database systems to evaluate its efficiency with the possibility of improving its functionality.

3. Natural Language Framework

As defined in standard dictionaries, a grammar is the study of the way the sentences of a language are constructed; morphology, syntax and semantics. Thus a grammar is the set of structural rules governing the composition of clauses, phrases, and words in any given language. NLP use parsing which is the process of analyzing a string of words in order to recover the structure of the sentence, either in natural or in computer languages, according to the rules of a formal grammar. The term parsing comes from Latin pars (noun), meaning part (of speech) [8]. Bottom-up parsers, such as the Chart parser or SRL parser, are suitable for analyzing natural language grammars as they eliminate the backtracking generated by Top-down recognizers optimizing the parsing process, but still present a combinatorial explosion problem.

The grammar defines categories of language (sentence, noun, verb, etc.). A grammatical category has some specific features, a feature is a property of items within the grammar of a language; a feature has a number of possible values which are normally mutually exclusive within a given category. Typical features encountered in grammatical categories include tense (e.g. present, past, etc.), number (singular and plural), and gender (masculine and feminine). In categorical grammars, categories have semantic distinctions; this is reflected in the contribution that each category has in the meaning of the sentence. But in generative grammars, the meaning is considered as separated from grammar; their categories define the distribution of syntactic elements [6].

3.1 Grammar for a subset of Spanish

The original system NatLIn, worked with a top-down parser in which grammar rules were structured within the code of the program, which made very difficult the adaptability of the system for new applications. This fact as well made very difficult the recovery of the original grammar rules. The semantic of the system is based in Montague Grammar [2] and makes use of complex symbol manipulation for transforming the Spanish question into first order logic using lambda calculus. Once the grammar was recovered it was incorporated to PyNatLIn system as an external construction for a Bottom-Up parser, making it easier the incorporation of new grammatical constructions, such as the rules for a better management of propositional phrases. The grammar is defined by a categorial grammar in the sense of Montague [2]. The semantics of the grammar is defined in a one-to-one correspondence with the syntactic rules in such a way that is possible to attain a compositional reconstruction of the meaning of the sentence. Some of the rules are presented in Fig. 1.

Syntactic Rules	Semantic Rules
1. $S \rightarrow NP\ VP$	$S' = NP'(VP')$
3. $NP \rightarrow Det\ Noun$	$NP' = Det'(Noun')$
4. $NP \rightarrow IP\ Noun$	$NP' = IP'(Noun')$
5. $VP \rightarrow Vt\ NP$	$VP' = \lambda Z.((NP'(Z))(VP')$
6. $Noun \rightarrow Noun\ Adj$	$Noun' = \lambda w.(Noun'(w)\ \&\ Adj'(w))$

7. $Noun \rightarrow Noun\ RP$	$Noun' = \lambda z.((Noun'(z))\ \&\ (RP'(z))$
8. $RP \rightarrow que\ VP$	$RP' = VP'$
9. $Det \rightarrow algúnos$	$Det' = some' = \lambda P.(\lambda Q.(exist(x, P(x) \& Q(x)))$
10. $IP' \rightarrow Qué$	$IP' = which' = \lambda P.(\lambda Q.(wh(x, P(x) \& Q(x)))$
11. $Adj \rightarrow Europeo$	$Adj' = 'europeo' = \lambda u.(europeo(u))$
12. $PN \rightarrow Australia$	$PN' = [australia]' = \lambda P.(P(australia))$
13. $PN \rightarrow El\_Mediterráneo$	$PN' = [mediterraneo]' = \lambda P.(P(mediterráneo))$
14. $Noun \rightarrow país$	$Noun' = 'country' = \lambda y.(pais(y))$
15. $Vt \rightarrow colindan$	$Vt' = 'borders'\ \lambda x.(\lambda y.(colinda(x, y)))$

Fig. 1

4. System Architecture

Spanish is a relatively inflected language, with a two-gender noun system and about fifty conjugated forms per verb, but with inflection of nouns, adjectives, and determiners limited to number and gender. Spanish syntax is considered right branching, meaning that subordinate or modifying constituents tend to be placed after their head words. Some examples of sentences analyzed by the system are the following:

a) ¿Que países colindan con el Mediterráneo?
(Which countries border with the Mediterranean?)
b) ¿Que países europeos colindan con aquellos países que colindan con el Mediterráneo?
(Which European countries border with any country which borders with the Mediterranean?)

To interpret the those sentences, we use and apply the rules from previous section to obtain the logical expressions:

a) $prg(z, colinda\_con(z, mediterráneo))$
b) $prg(y, exists(x, país(x)\ \&\ colinda\_con(x, mediterráneo)\ \&\ colinda\_con(x, y)))$

The following figure 1, shown the way in which PyNatLIn interprets Spanish queries:

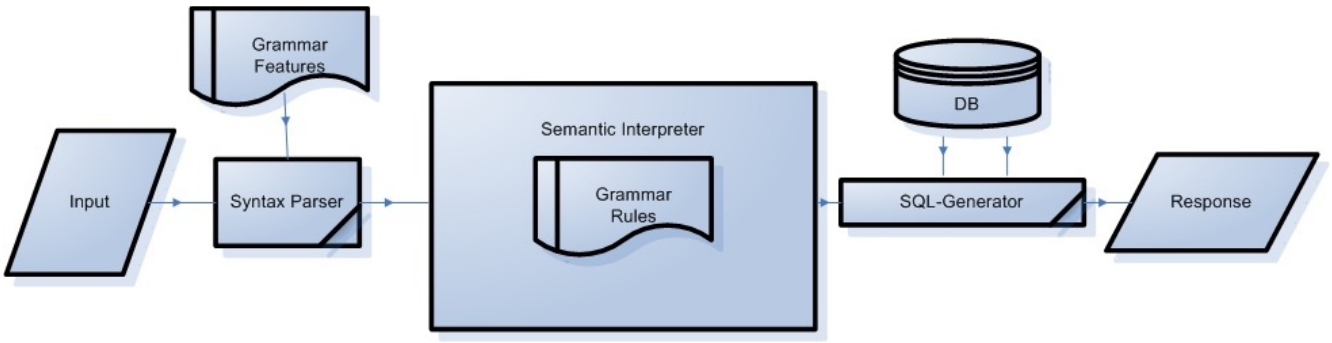

Fig. 2

The system receives the Spanish question and is analyzed by a natural language processor that uses syntax and semantic rules that provide an interpretation of the sentence. The meaning of the sentence is expressed as a formula in first order logic. The formula then is translated into a SQL expression that can be evaluated in a commercial database system.

5. Conclusion and future work

This paper proposes a Spanish Query Answering System, PyNatLIn, which explores Natural Language Processing in Spanish, the model and design framework are based on the original system NatLIn [1,3,5] built in Prolog. Its implementation in Python is based on the NLTK library [4] that provides easy-to-use suite of text processing routines for classification, tokenization, stemming, tagging, parsing, and semantic reasoning for applications of Natural Language Processing. The PyNatLIn is a good example of a Natural Language Processing system for interpreting Spanish queries to a relational database formation of complex SQL queries to test and improve the efficiency and accuracy of the proposed architecture. In future we will extend the implementation for Spanish Query Answering System, implementing an extendable grammar interface to cover different types of Spanish queries and an interface for allowing the adaptation of the system to new domains.

6. References

[1] R. Dávila, Osorio M., Quintero M.: Interfaz en Español para Recuperación de Información en una Base de Datos Geográfica: Proceedings of the CONIELECOMP 2002, Acapulco, Mexico, February (2002).

[2] R. Montague: The Proper Treatment of Quantification in Ordinary English: In: R. H. Thomason, editor: *Formal Philosophy. Selected Papers of Richard Montague*: Yale University Press, New Haven (1974).

[3] R. Dávila-Pérez: Access in Spanish to a Predicate Calculus Database: M. Sc. Dissertation: University of Essex, England (1986).

[4] Natural Language Tool Kit http://www.nltk.org

[5] Dávila-Pérez R., 2002, *Analisis and Interpretaton of Negative Queries in Spanish to a Geographical Databasel,* Memorias del MICAI/TAINA 2002, Avances en Inteligencia Artificial, Mérida, México (Abril 2002).

[6] Joan Bybee "Irrealis" as a Grammatical Category. Anthropological Linguistics, Vol. 40, No. 2 (Summer, 1998), pp. 257-271

[7] Pereira F. and Warren D., An Efficient Easily Adaptable System for Interpreting Natural Language Queries, American Journal of Computational Linguistics, Volume 8, September 3-4, July-December 1982.

[8] Definition parser [online] http://dictionary.reference.com/browse/parse

[9] V. Dahl: Translating Spanish into Logic through Logic. *American Journal of Computational Linguistics*, 13:149-164, (1981).