INQUER: A WordNet-based Question-Answering Application

Catarina Ribeiro, Ricardo Santos, João Correia, Rui Pedro Chaves and Palmira Marrafa

Universidade de Lisboa, CLUL
CLG – Computation of Lexical and Grammatical Knowledge Research Group
Av. Prof. Gama Pinto, 2
1649-003 Lisboa, Portugal
{catarina.ribeiro,ricardo.santos}@clul.ul.pt, jcor@net.sapo.pt, rui.chaves@clul.ul.pt, Palmira.Marrafa@netcabo.pt

Abstract

The growing use of computer technologies as well as the amount of available information has posed new challenges to human-machine interaction. Applying natural language processing techniques to information systems has brought advantages to this emerging field. Question-answering (QA) systems have been developed in order to ease the information access process. However, most of today's QA systems retrieve a collection of documents whose contents may fulfill the answer to user's queries. Such systems usually consider unstructured information sources, namely Web or textual documents. Adopting a different approach, INQUER, the QA system described in this paper, makes use of a structured knowledge base: WordNet.PT (Portuguese WordNet). This system provides direct natural language answers to user's questions by applying inference and information extraction mechanisms that interact with WordNet.PT (Portuguese WordNet). Although INQUER deals specifically with Portuguese, the theoretical approach that underlies this application is language independent.

Keywords: Natural Language Question-Answering, WordNet, Semantic Inference.

1. Introduction

The interest on Question-Answering systems has been growing worldwide over the last few years, due to the progress of technology in what concerns Knowledge Engineering. Current state of the art QA systems involve information retrieval from a large collection of text and are usually Web-based.

Information retrieval plays a crucial role in this kind of systems due to their need of broad coverage. These systems are able to answer questions that have brief phrasal answers ('factoids'), by identifying and extracting the answer from texts, e.g. LAMP (Zhang & Lee, 2003).

Most of today's Natural Language Question-Answering systems should rather be called answer extraction systems since they are only able to find answers explicitly included in the source texts – e.g. AnswerBus (Zheng, 2002), IONaut (Abney at al., 2000). However, some systems proposed in the last few years combine natural language techniques with information retrieval ones in order to deal with more complex questions. In addition, some of these approaches consider both basic lexical semantic relations (e.g. Litkowski, 2000) and WordNet structure (e.g. Litkowski, 2001; Hermjakob et al., 2002) as well as statistical methods to score candidate answers.

The QA system described in this paper – INQUER – implements a different approach: it does not rely on the probabilistic extraction of sections from large collections of text, rather it uses structured information from a linguistic ontology. Our system allows users to interact with WordNet.PT through unrestricted natural language questions, i.e. it is not constrained to template queries. A syntactic-semantic analyzer is applied not only to analyze the question but also to build a semantic representation (first-order logical form). Inference and information extraction mechanisms are then applied (on the fly) to extract the relevant information from the Knowledge Base (KB). In a last step, a natural language answer is generated based both on the information extracted and on the question representation. INQUER system takes advantage of wordnet internal architecture and provides a user-friendly interaction.

In section 2 a brief overview of the Portuguese wordnet is presented. A description of each module of our QA system is given in section 3 and an online demo of INQUER is described in section 4. Finally, section 5 contains concluding remarks and future work.

2. WordNet.PT

WordNet.PT is being developed within the general framework of EuroWordNet (Vossen, 1999). EuroWordNet is a multilingual database with individual wordnets for several different European languages interrelated by an Inter-Lingual-Index. The individual wordnets are basically structured along the same lines of the Princeton WordNet (Miller et al., 1990; Fellbaum, 1998).

A wordnet is a lexical-conceptual database whose basic units are lexicalized concepts (single words or complex sequences) grouped in a 'synset' (set of synonyms). The meaning of a lexical unit is derived from the lexical-semantic relations it establishes with other members of the same synset as well as with other synsets. Synonym is, then, the most basic relation in wordnets.

The hyponym/hyperonym relation is the most fundamental structuring relation in wordnets. It can be informally defined as follows: A is a hyponym of B (B hyperonym of A) iff A is a kind of B and B is not a kind of A.

The part/whole relation is another major relation coded in wordnets. WordNet.PT distinguishes five part/whole relation subtypes and differentiates between canonical and non canonical part/whole relation.

In WordNet.PT there is a gloss – informal definition – associated with each concept that specifies additional information.

3. INQUER System

INQUER system uses the Portuguese wordnet both as a lexical database – to analyze and generate natural language questions – and as a semantic knowledge base – to obtain answers via an inference engine.

All the components of this system were implemented using ProLog language. The INQUER system architecture is divided in three major modules: (i) a syntactic and semantic parser; (ii) an inference and information extraction engine and (iii) a natural language answer generator.

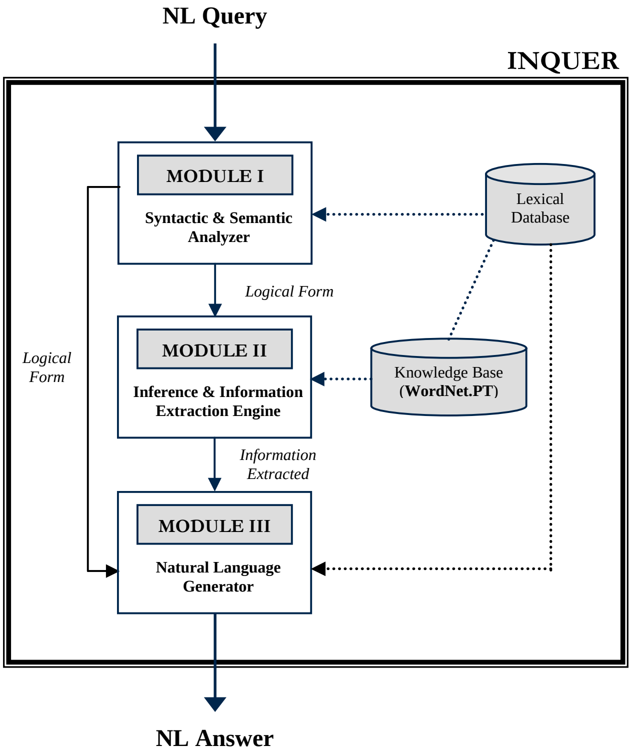
NL Query
INQUER
MODULE I
Syntactic & Semantic Analyzer
Lexical Database
Logical Form
Logical Form
MODULE II
Inference & Information Extraction Engine
Knowledge Base (WordNet.PT)
Information Extracted
MODULE III
Natural Language Generator
NL Answer

Figure 1: General architecture of INQUER system

Figure 1 describes the general architecture of INQUER system. The question-answering process involves three sequential major modules and starts with the input of a natural language query. The first module converts the input into a first-order logic form. To do so, the syntactic-semantic analyzer interacts with the lexical database. The logical form is then fed into the inference and information extraction engine that browses the KB selecting the required information. The output of the later module and the logical form are both fed into the natural language generator in order to produce a final answer.

A more detailed description of each module is provided in the next sub-sections.

3.1. Module I: Syntactic and Semantic Analyzer

The main task of first module is to develop a Portuguese grammar fragment with a phrase-structure rule backbone and a HPSG-based subcategorization mechanism. This formalism conveys important generalizations since it deals with different kinds of linguistic phenomena in a single rule. The grammar rules are subdivided into three types: meta-lexical entries (that enrich the lexical entries with a first-order logic representation), lexical rules (that capture linguistic phenomena such as *gaps* and expletive null subject) and phrase structure rules (that encode syntactic and semantic rules). The analyzer makes use of a tabular bottom-up parser adapted from Gazdar & Mellish (1989).

On the lexical database, each entry contains its specific semantic and grammatical attributes. The nature of the hyperonymy relation legitimates the straightforward application of transitive inheritance mechanisms that recursively assign missing attributes from the direct hyperonym. The system is also sensitive to multiword units and lexical and syntactical ambiguities.

This module internal architecture is presented in Figure 2.

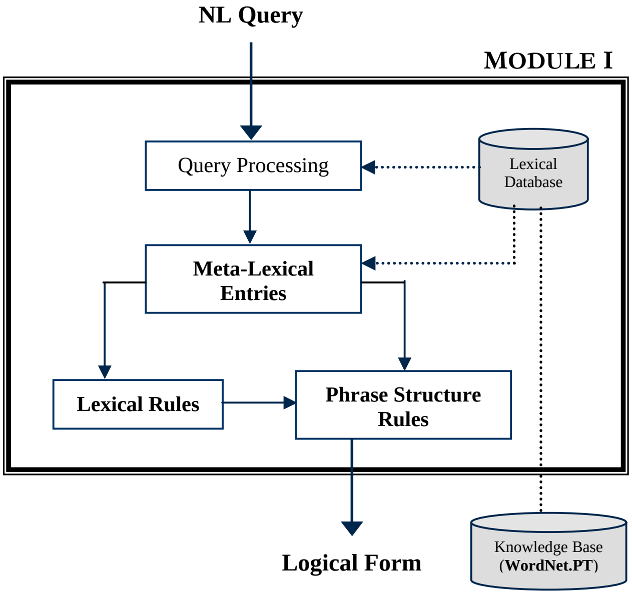

Figure 2: Internal architecture of Module I

For each introduced question the system uses the lexical database to recognize its basic units. At this stage multiwords are also identified. The next step is to assign each word form a syntactic and semantic representation using meta-lexical entries. However, if special linguistic phenomena are at stake, lexical rules are applied. Using phrase structure rules, the grammar analyses the sentence and builds a semantic representation. This is a compositional process based on higher-order lambda terms implemented in Prolog (cf. Pereira & Schieber, 1987). The semantic representation of the sentence is a first-order logical form which is fed into the second module (Inference and information extraction engine).

If the user introduces, for example, a question such as (1):

1) “Quais são os carnívoros que nadam e que não são
‘which are the carnivores that swim and that not are
mamíferos?”
mammals?’

the output of module I would be (2):

2) which(A): (∀B (carnivore(B) ∧ swim(B) ∧ ¬(∃C (mammal(C) ∧ be(B,C))) → be(A,B)))[1].

3.2. Module II: Inference and Information Extraction Engine

Module II implements a first-order logic model checker based on Blackburn & Bos (1999) for yes/no questions (e.g. “Apples have K vitamins?”) and searching

[1] For ease of exposition the logical form has been translated.

algorithms for information retrieval for wh-questions (e.g. “What kind of animals have gills?”) and definition questions (e.g. “What is a siamese cat?”).

The internal architecture of this module is described in Figure 3.

The first step of the process is the identification of the type of question introduced by the user (through the analysis of the head of the logical form) in order to decide which algorithm to apply. If it is the case of a yes/no question, a first-order logic model checker and a search algorithm are applied. The former tries to satisfy the logical form considering a closed-world assumption, i.e., all the non-satisfiable logical forms are considered to be false. Since the system handles first-order formulas, it is sensible to different kinds of quantification (universal and existential – cf. Sit & Kolackovsky, 1998) so, for instance, it may decide that it is true that “Some birds that do not fly swim” but is not case the that “Every bird that does not fly swims”.

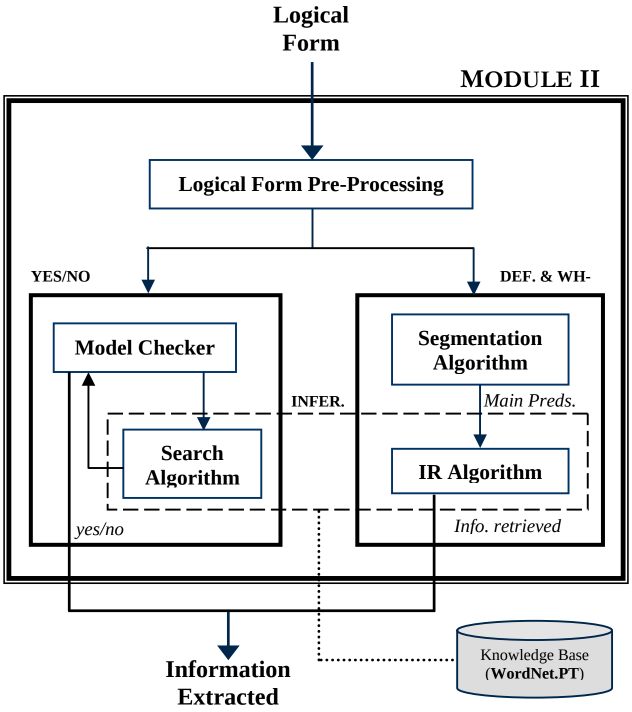

Figure 3: Internal architecture of Module II

The model checker makes use of a search algorithm that interacts directly with the KB. Such algorithm applies mechanisms that are capable of infer implicit information.

On the other hand, if it is the case of a definition or a wh-question, the logical form is decomposed in its main predicates. An information retrieval (IR) algorithm is then applied in order to extract sets of concepts (or a gloss) that make the logical form satisfiable on the knowledge base. This group of selected concepts will be pruned so that only the most general hyperonyms are outputted. For example, instead of “dog, wolf, fox, etc.”, the system would retrieve “canine”.

The output of module II – yes/no or sets of concepts depending on the type of question – is, together with the initial logical form, fed into the natural language generator. Considering a logical form such as (2), the extracted information would be (3):

3) “crocodilo, jacaré, tubarão-martelo, tubarão-tigre,
‘crocodile, alligator, hammerhead shark, tiger shark,
tubarão branco, piranha”
white shark, piranha’.

3.3. Module III: Natural Language Answer Generator

The main goal of this module is to improve the quality of the answer provided to the user, since it returns the information in natural language sentence format. It is necessary to consider the human behavior to augment the usability and the acceptance of an information system (see Baecker & Buxton, 1987).

Module III implements a mechanism - the Combining Algorithm - which uses the output of the previous two modules (logical form and information extracted from WordNet.PT) and returns a natural language answer in Portuguese.

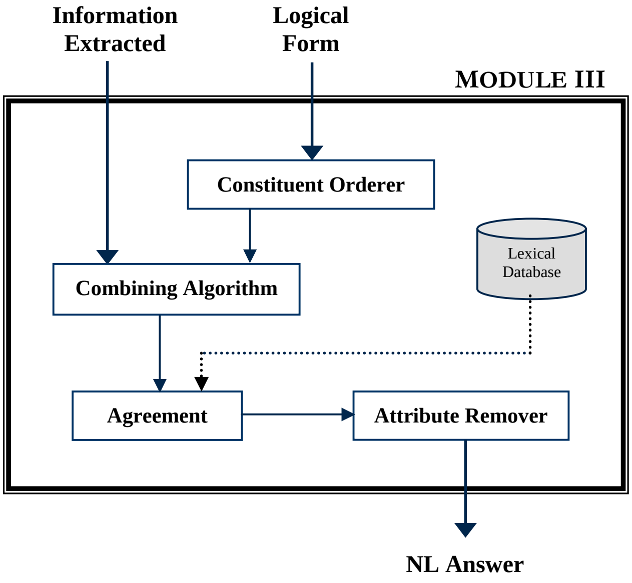

Figure 4: Internal architecture of Module III

If the introduced question is not valid the answer produced by the generator is “Expressão Inválida” (*Invalid Expression*). If it is the case of a “how many” question, the output of module II is the count of concepts that satisfy the logical form. The answer produced by the generator is the corresponding cardinal number.

On other situations, the Combining Algorithm defines the structure of the natural language answer based on the logical form of the question. In order to generate an answer template, the mechanism uses an auxiliary algorithm – the Constituent Orderer – that structures a phrase or a sentence into SVO order. This algorithm loosely follows the method in Shieber et al. (1989), building complex structures compositionally. The next step is to apply agreement rules and insert, if needed, new lexical material to generate a grammatical answer. Finally, all the attributes needed to make the agreement and other unnecessary elements are removed.

For example, if the input of module III is a logical form such as (2) and a set of concepts obtained from this logical form (3), the generated answer would be (4):

4) "Os carnívoros que nadam e não são mamíferos são:
'the carnivores that swim and not are mammals are:
crocodilo, jacaré, tubarão-martelo, tubarão-tigre, tubarão
crocodile, alligator, hammerhead shark, tiger shark, white
branco e piranha"
shark and piranha'.

4. Online Demo

The INQUER system has a website where a demo version is available at http://www.clul.ul.pt/clg/inquer. In this demo only the "food" and "living beings" domains are considered. The lexical database contains 57 (10 transitive and 47 intransitive) verbal forms and 3723 common nouns.

The user can write a question (in Portuguese) or pick an example from a list. The sentences accepted in this version are definition questions (return a gloss), yes/no questions (return yes/no) and wh- questions (return a value or a list of concepts). Several linguistic phenomena were considered such as prepositional phrase attachment, negation, coordination and relative clauses. If the question introduced by the user is ambiguous – either lexically or syntactically – different answers can be obtained (one per interpretation).

After the submission of the question, a response webpage is showed to the user. Along with the final answer, additional information is provided such as the run time for each module and the logical form.

5. Conclusions and Future Work

A large-scale linguistic database such as WordNet.PT opens quite challenging possibilities within several other domains of Natural Language Processing and Language Technologies. In the work reported here WordNet.PT is explored as a semantic knowledge base and as a lexical database of a Question-Answering system. The system described in this paper is flexible enough to be easily adapted to other lexical-relational databases.

The INQUER system uses deep syntactic and semantic analysis to produce a first-order formula, which is then fed into an inference engine that uses WordNet.PT as a semantic knowledge base. The system allows users to formulate yes/no, definition and wh-questions in Portuguese and to obtain explicit and inferred information from the database. The system also includes generation rules that allow users to obtain an answer in natural language instead of in a logical formula.

In the future, the inference engine should be extended to other semantic relations and new linguistic phenomena should be introduced. Since the system is sensitive to syntactic and lexical ambiguities, an answer per interpretation is shown and a probabilistic analysis to order the solutions is being considered.

Furthermore, in order to deal with more complex questions, such as those involving choices or comparisons, the internal analyses of glosses is being considered.

References

Abney, S.; Collins, M. & Singhal, A. (2000). Answer Extraction. In Proceedings of Applied Natural Language Processing Conference (ANLP).

Baecker, R. M. & Buxton, W. A. S. (eds.) (1987). Readings in Human-Computer Interaction: A Multidisciplinary Approach. San Mateo, CA: Morgan Kaufmann Publishers.

Blackburn, P. & Bos, J. (1999). Representation and Inference for Natural Language, a First Course in Computational Semantics. Stanford: CSLI Press (forthcoming).

Fellbaum, C. (1998). A Semantic Network of English: The Mother of All WordNets. In P. Vossen (ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.

Gazdar, G. & Mellish, C. (1989). Natural Language Processing in Prolog. Workingham: Addison Wesley.

Hermjakob, U.; Hovy, E. H. & Lin, C. (2002). Knowledge-Based Question Answering. In Proceedings of the 6th World Multiconference on Systems, Cybernetics and Informatics (SCI-2002), Orlando, FL, U.S.A., July 14-18, 2002.

Litkowski, K. C. (2000). Question-Answering Using Semantic Relation Triples. In Voorhees, E. M. & Harman D. K. (eds.) Information Technology: The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246 (pp. 349--356). Gaithersburg, MD: National Institute of Standards and Technology.

Litkowski, K. C. (2001). Syntactic Clues and Lexical Resources in Question-Answering. In Voorhees, E. M. & Harman D. K. (eds.) Information Technology: The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-249 (pp. 157--166). Gaithersburg, MD: National Institute of Standards and Technology.

Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. In International Journal of Lexicography, Vol.3, Nº 4 (pp.235—244).

Pereira, F & Shieber, S. (1987). Prolog and Natural-Language Analysis. In CSLI Lecture Notes, 10. Stanford, California: Cambridge University Press.

Shieber, S.; Noord, G.; Moore, R. & Pereira, F. (1989). A Semantic-Head-Driven Generation Algorithm for Unification-Based Formalisms. Artificial Intelligence Center, Menlo Park, USA.

Sit, E. & Kolackovsky, P. (1998). Artificial Intelligence – Inference in First-Order Logic – Part 2. University of Calgary.

Vossen, P. (1999). EuroWordNet General Document. University of Amsterdam.

Zhang, D. & Lee, W. (2003). A Web-based Question-Answering System. In Proceedings of the SMA Annual Symposium 2003, NUS, Singapore.

Zheng, Z. (2002). AnswerBus Question-Answering System. In Proceedings of Human Language Technology Conference (HLT2002), San Diego, CA.