The University of Évora approach to QA@CLEF-2004

Paulo Quaresma and Luís Quintano and Irene Rodrigues and José Saias and Pedro Salgueiro
Departamento de Informática, Universidade de Évora, Portugal
pq,ljcq,ipr,jsaias,ps@di.uevora.pt

Abstract

The approach followed by the University of Évora team in order to build a system able to participate in the QA-CLEF task is described.

The system is based in two steps: for each question, a first information retrieval task selects a set of potentially relevant documents; then, each of these documents is analysed trying to obtain their semantic representation and the answer to the initial query.

The proposed approach was applied to the test set of the QA-CLEF for the Portuguese language and the obtained results were quite interesting and motivating and allowed the identification of strong and weak characteristics of the system.

1 Introduction

Question answering systems are an important topic of research in the natural language processing field and much work has been done by many researchers in the last years. Several international conferences have special tracks for analysing this topic, namely, the TREC – Text REtrieval Conference (http://trec.nist.gov) or the CLEF – Cross Language Evaluation Forum (http://www.clef-campaign.org).

For the 2004 campaign, CLEF has added the Portuguese language as a possible language for the queries and for the target documents.

In the last years, the Informatics Department of the University of Évora has been working in the natural language processing field, namely trying to develop specialised tools for the Portuguese language.

This paper describes the University of Évora approach to the question answering task for the Portuguese language of CLEF 2004. The collection of target documents is the set of news published by the Portuguese newspaper "Público" during 1994 and 1995. Questions (200) can be factoids or definitions and some of them may have no answer in the target set of documents.

The proposed system is based in two steps:

- For each question, a first information retrieval task selects a set of potentially relevant documents.
- Then, for each of these documents that where analysed in the preparatory phase trying to extract the facts they conveyed, the user query is interpreted on each selected text knowledge base. When an answer to the query is obtained, the process stops and the system outputs the answer and then identifies the document from where the answer was obtained.

Our question answer system needs to have a preliminary information retrieval task, defining a smaller set of potentially relevant documents due to computational complexity problems. In fact, our main approach is to deeply analyse the set of documents and to obtain a partial semantic representation of their content. Then, each query would be transformed into its semantic form

and an inference process would try to obtain the answer to the query. However, this approach showed many scalability problems due to the large number of documents and associated data and it was necessary to strongly reduce its cardinality.

Section 2 describes the preparatory phase where the set of documents are preprocessed in order to build the IR indexes and the knowledge base for each text.

Section 3 describes the proposed architecture for the question-answer system and section 4 describes each of the architecture modules. A preliminary evaluation is presented in section 5 and some conclusions and future work is discussed in section 6.

2 Pre-processing the set of target documents

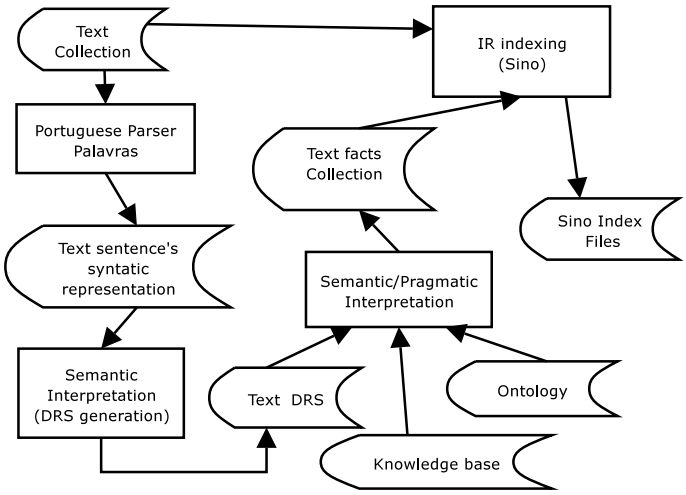

Figure 1: Texts preprocessing

There exists an important pre-processing phase of the target collection of documents in order to obtain the input data necessary for our question answer system.

In this phase two main tasks are done:

- *Semantic/Pragmatic Interpretation* – gives rise to a set of knowledge bases, *Text facts collection*, where each knowledge base has the facts conveyed by each text.
- *Information retrieval indexing* – creates the files that index the full set of documents with a reference to the knowledge base associated to each document, *Sino Index Files*.

The other tasks of this phase are:

- Portuguese Parser - each text of the collection is analysed by the Portuguese parser PALAVRAS [3] developed in the context of the VISL[1] project in the *Institute of Language and Communication* of the *University of Southern Denmark*. The output of the parser is a file with the syntactic analysis of each text.

  We've chosen to keep the first syntactic analysis for each sentence; this option is one of our sources of problems.

- Semantic Interpretation – each syntactic structure is rewritten into a First-Order Logic expression. The technique used for this analysis is based on DRS's (Discourse Representation Structures)[6].

[1] Visual Interactive Syntax Learning

This technique identifies triggering syntactic configurations on the global sentence structure, which activates the rewriting rules. We always rewrite the pp's by the relation 'rel(prep,A,B)' postponing its interpretation to the semantic pragmatic module.

The semantic representation of sentence is a DRS build with two lists, one with the new sentence rewritten and the other with the sentence discourse referents.

One of the most important requirements of the proposed QA system is to have a knowledge base of facts inferred from the analysis of the set of target documents and an ontology containing the concepts referred in the documents.

- Ontology – From the output of the DRS's generation and from an existent top ontology of concepts, a new ontology containing the concepts referred in the documents was created [10, 9].

  This step showed to be very problematic, due to the large number of concepts referred in the documents and to the complexity and difficulty of finding correct relations between them.

  The obtained ontology was created in the OWL (Ontology Web Language) format and in a logic programming framework, ISCO [1, 2], which allows the integration of Prolog-like inference mechanisms with classes and inheritance, and constraint solving algorithms.

- Knowledge base – From this ontology and from each sentence semantic representation we could obtain the interpretation of each text sentence that will give rise to a set of facts to add into out knowledge base [7].

  However, this task showed to be very complex in its computational time and space and the obtained knowledge base was very large and it created many problems to the inference processes.

  Accordingly, it was decided to first decrease the set of relevant documents to each query (via IR techniques) and, then, to create a set smaller knowledge base.

  The knowledge base referred in figure 1 was build with a set of facts extracted from the target text collection and with rules and facts that we import from other applications.

2.1 Semantic/Pragmatic Interpretation of text sentences

In order to obtain the set of facts of each text sentence we need to use the ontology in order to obtain the meaning of each text sentence.

The semantic/pragmatic module receives the sentence rewritten (into a First Order Logic form) and tries to interpret it in the context of the document database information (ontology).

In order to achieve this behaviour the system tries to find the best explanations for the sentence logic form to be true in the knowledge base for the semantic/pragmatic interpretation. This strategy for interpretation is known as "interpretation as abduction" [5].

The knowledge base for the semantic/pragmatic interpretation is built from the Ontology. The inference in this knowledge base uses abduction, restrictions (GNU Prolog Finite Domain (FD) constraint solver).

The knowledge base rules contains the information for the interpretation of each term in the sentence logic form as a prolog term.

As an example consider the sentence:

"O gato do João comeu o rato do Manuel/John's cat ate Manuel's mouse."

is transformed into a DRS-like term showing the 4 referents and their relations:

```
drs([def-A-m-s, def-B-m-s,
     def-C-m-s, def-D-m-s],
    [cat(A), rel(of,A,B),
     name(B,'João'), comer(A,C),
     mouse(C), rel(of,C,D),
     name(D,'Manuel')]).
```

The semantic interpretation module using the ontology will rewrite this DRS into:

```
drs([def-A-m-s, def-B-m-s,
     def-C-m-s, def-D-m-s],
    [cat(A), owns(B,A), person(B),
     name(B,'João'), eats(A,C),
     mouse(C), owns(D,C), person(D),
     name(D,'Manuel')]).
```

The interpretation of *rel(of,A,B)* as owns(A,B) is possible due to the existence of the relation *owns* that relates persons and animals.

Other important step in this task is to create new individuals (new identifiers) for discourse referents when they are not instantiated during the interpretation.

This last step is a source of problems for our QA-system since it is possible to have different identifiers for the same individual if this task fails to identify the sentences entities. The opposite can also happen, this task may unify individuals that are different.

The option of building a knowledge base with the facts extracted for each documents helps us to deal with this problems, there are fewer entities.

A problem that we still have to solve is the way how we choose the best meaning for a sentence.

2.2 Information retrieval indexing

SINO [8, 4], originally from the Australasian Legal Information Institute, was used to index the full set of documents. It allows the creation of inverted index files and in the new version it uses information specific to the Portuguese language: stop words and lemmatization. In fact, SINO was extended to use a set of Portuguese stop words (such as, articles, pronouns, prepositions) and to transform each word in its lemma (using the Portuguese lexicon POLARIS).

Documents are indexed by a specialized search engine for the Portuguese language – SINO [8, 4] – and an information retrieval system for this collection is built. As it will be described in more detail in the next section, the information retrieval system will be used, for each query, to decrease the cardinality of the target set of documents.

3 Architecture

The architecture is composed by several independent modules. Figure 2 shows a graphical view of their relations.

In the next sub-sections a brief description of each module is presented.

3.1 Query processing

Each query is processed using the same natural language tools used to analyse the full set of documents, The Portuguese syntactic Parser Palavras and the DRS's generation. After obtaining the Query DRS there are two task that are performed concurrently:

- The Semantic Pragmatic Interpretation of the query. The query semantic representation is obtained taking into account the ontology of concepts and a knowledge base with some general world knowledge.
- The query preprocessing and the interrogation of the IR system.

  After obtaining the query DRS, it is transformed into a search term to the IR engine – SINO. This step is needed because it was computationally impossible to handle inferences over the complete knowledge base created in the pre-processing phase of the documents. So, we use an information retrieval system to obtain a set of relevant documents to make inferences only over the knowledge base created with the information conveyed by these documents.

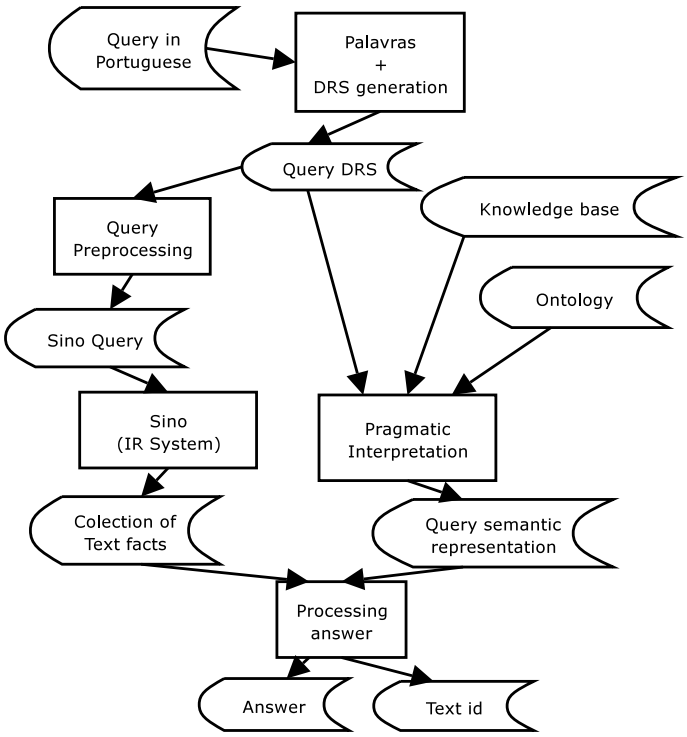

Figure 2: QA System's architecture

The IR queries are created from the semantic representation, DRS, of each query and their structure will be described in the next section.

Using the IR queries, the search engine obtains an ordered set of relevant documents. This set is used to create a smaller knowledge base of DRS containing only the information conveyed by these relevant documents.

In this way it is possible to strongly decrease the complexity of the knowledge base and it is possible to handle inferences over it.

Finally the *Process answer* task receives the set of relevant knowledge bases where the query semantic/pragmatic representation should be evaluated and tries to infer the answer to the query.

The inference process is based in a logic programming framework, ISCO [1, 2], which allows the integration of Prolog-like inference mechanisms with classes and inheritance, and constraint solving algorithms.

4 Modules

In this section a more detailed description of the most relevant modules of the architecture is presented.

4.1 Natural Language Query Processing

This module tasks: Palavras+Drs generation and the Pragmatic Interpretation follow an approach similar with the semantic-pragmatic interpretation of the documents sentences and it uses the same natural language tools: the PALAVRAS parser, the DRS generation and semantic-pragmatic interpreter.

After the DRS generation we are able to identify the referents which are the focus of each query and the kind of query performed.

For instance, the query: "Quem comeu o rato do Manuel/Who ate Manuel's rat?" is transformed into the DRS-like term:

```
drs([who-A-X-Y, def-B-m-s, def-C-m-s],
    [eat(A,B),
     mouse(B), rel(de,B,C),
     name(C,'Manuel')]).
```

After obtaining the query DRS, the semantic-pragmatic interpretation using the ontology of concepts created in the pre-processing phase gives rise to the final query representation:
For the above example it will be:

```
drs([who-A-X-Y, def-B-m-s, def-C-m-s],
    [eat(A,B),
     mouse(B), owns(C,B),person(C),
     name(C,'Manuel')]).
```

This final query representation will be evaluated in each knowledge base selected by the last sino query.

4.2 Query preprocessing

After obtaining the query DRS, it is transformed into a search term to the IR engine – SINO.

The approach followed was to create three IR query terms for each natural language query and to order the set of documents retrieved. The overall idea is to create a very restrictive query, a very general one, and one in the middle. The created IR queries are boolean ones and they are obtained from the DRS of each query:

```
‘‘Em que cidade se encontra a
  prisão de San Vittore?’’

cidade AND encontrar AND prisão
AND (San  AND Vittore)

cidade AND (encontrar OR
prisão OR (San  AND Vittore))

cidade OR encontrar OR prisão
OR (San  AND Vittore)
```

The first query is the more specific, obtained from the boolean AND of each term; the second query is obtained from the boolean AND of the head of the query with the OR of the other terms; and the third query is the more general, obtained from the OR of each term.

From the ordered set of documents, the first 50 are selected and they are the basis for the creation of each query-related knowledge base.

4.3 Sino - Relevant documents DRS extraction

This module receives the three IR queries and it retrieves the correspondent relevant documents. As it was already described, the IR engine used was an extension of the SINO engine from the AustLII institute.

SINO retrieves the relevant documents (using the boolean operators) and it orders the selection using a ranking function. This ranking function gives higher priority to documents with more word hits or with hits in the title. It is important to point out that first documents are ordered accordingly with the kind of query: first the documents retrieved from the more specific query and

with last priority the documents retrieved from the more general one. Inside each set of retrieved documents, the order is obtained through the SINO ranking function.

After having a ordered list of relevant documents, the first 50 were selected as the basis to create a knowledge base of facts relevant to the query. The reason why the first 50 were chosen is related with the goal of reducing the computational complexity and assuming a good performance of the SINO engine.

4.4 Answer inference process

This module is the responsible for finding the correct, exact answer to each query. It receives the semantic-pragmatic interpretation of each query (in a DRS-like format) and a logic-programming based knowledge base built from the set of the most relevant 50 documents of each query.

The inference process is done via the use of the Prolog resolution algorithm, which tries to unify the referent in the query with facts extracted from the documents. This unification takes into account the information associated with the referents, such as, genre and number. Moreover, the inference process uses the "kind of" question, such as, where/when/who, to identify the feature that is queried about. For instance, if the query is about a place of a specific entity, "Em que cidade se encontra a prisão de San Vittore?", the system tries to find a feature of that entity that is a place and it is not referred in the query.

As it can be seen from this description, the proposed system relies on the quality of the inferred ontology and in a good semantic-pragmatic interpretation of sentences and queries.

As a consequence of our approach, the system has always a confidence value for each answer of 1: if it finds an answer, then it is sure about it!

5 Evaluation

The proposed system was applied to the set of 200 questions (in fact they were 199, because one question was not considered by the judges).

It obtained 47 correct answers, 18 inexact and 134 wrong with an overall accuracy of correct answers of 23.62% and a confidence-weighted score of 0.21619. If the accuracy is calculated over the correct and the inexact answers, then its value is 32.66%.

We believe these are quite interesting results, that show the potential of the proposed approach. However, they also show the main problem of the system: it gave 127 "nil" answers and only 9 of them were correct.

The most important question now is: what happened in the 118 questions that had no answer from the system?

A preliminary evaluation showed that there were two main causes of problems:

- the information retrieval system
- the ontology

The information retrieval system, which was used to decrease the complexity of the knowledge base, quite often was not able to find the relevant documents. In fact, this problem can be clearly seen in the results of the IR task of CLEF'04, in which SINO showed very low recall values. This problem can be overcome by changing the SINO queries to better ones or by solving the complexity problems that made impossible the construction of a large, unique knowledge base. We intend to explore both possibilities.

The second problem was the quality of the ontology. In fact, the inference process relies heavily on the ontology. For instance, it is important to know what are places, persons, dates, synonyms. In the example presented previously, if the ontology does not have information relating "cidade" with the class of "places", then the system would not be able to answer the query. We will also continue to develop new strategies for constructing and merging ontologies.

6 Conclusions and Future Work

This proposal represents a first approach for a question answering system for the Portuguese language.

Our system uses natural language processing techniques to create a knowledge base from the information conveyed by the target documents. Queries are analysed by NLP tools and inferences are done over the knowledge base trying to find a correct answer. The inference process is done using a logic programming framework and the Prolog resolution.

The initial idea of creating a unique, large knowledge base with the facts extracted from all the documents was not feasible due to computational complexity problems. These problems led to the creation of an IR pre-analysis of the queries to decrease the complexity of the knowledge base. However, the IR engine showed some recall problems and lead to the incapacity of the QA system to answer many queries.

The ontology used was also a major problem and it was the origin of many other wrong answers. As the QA@CLEF task uses general domain documents, this is a very complex problem: how to obtain a good general purpose ontology?

As future work, we intend to try to develop new strategies for (partially) overcome these problems. Working with new implementation strategies it may be possible to have an unique knowledge base and using existent ontologies and Wordnets may improve the quality of the final ontology.

Finally, we also intend to explore the problem of inter-sentence anaphoric references and to be able to identify the correct referents in the documents.

References

[1] Salvador Abreu. Isco: A practical language for heterogeneous information system construction. In *Proceedings of INAP'01*, Tokyo, Japan, October 2001. INAP.

[2] Salvador Abreu, Paulo Quaresma, Luis Quintano, and Irene Rodrigues. A dialogue manager for accessing databases. In *13th European-Japanese Conference on Information Modelling and Knowledge Bases*, pages 213–224, Kitakyushu, Japan, June 2003. Kyushu Institute of Technology. To be published by IOS Press.

[3] Eckhard Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[4] G. Greenleaf, A. Mowbray, and G. King. Law on the net via austlii - 14 m hypertext links can't be right? In *In Information Online and On Disk'97 Conference, Sydney*, 1997.

[5] Jerry Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. Interpretation as abduction. Technical Report SRI Technical Note 499, 333 Ravenswood Ave., Menlo Park, CA 94025, 1990.

[6] Hans Kamp and Uwe Reyle. *From Discourse to Logic:An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: D. Reidel, 1993.

[7] Paulo Quaresma and Irene Pimenta Rodrigues. A natural language interface for information retrieval on semantic web documents. In E. Menasalvas, J. Segovia, and P. Szczepaniak, editors, *AWIC'2003 - Atlantic Web Intelligence Conference*, Lecture Notes in Artificial Intelligence LNCS/LNAI 2663, pages 142–154, Madrid, Spain, May 2003. Springer-Verlag.

[8] Paulo Quaresma and Irene Pimenta Rodrigues. PGR: Portuguese attorney general's office decisions on the web. In Bartenstein, Geske, Hannebauer, and Yoshie, editors, *Web-Knowledge Management and Decision Support*, Lecture Notes in Artificial Intelligence LNCS/LNAI 2543, pages 51–61. Springer-Verlag, 2003.

[9] José Saias. Uma metodologia para a construção automática de ontologias e a sua aplicação em sistemas de recuperação de informação – a methodology for the automatic creation of ontologies and its application in information retrieval systems. Master's thesis, University of Évora, Portugal, 2003. In Portuguese.

[10] José Saias and Paulo Quaresma. Using nlp techniques to create legal ontologies in a logic programming based web information retrieval system. In *Workshop on Legal Ontologies and Web based legal information management of the 9th International Conference on Artificial Intelligence and Law*, Edinburgh, Scotland, June 2003.