A question-answering system for Portuguese juridical documents

Paulo Quaresma
Departamento de Informática
Universidade de Évora
7000-671 Évora, Portugal
pq@di.uevora.pt

Irene Rodrigues
Departamento de Informática
Universidade de Évora
7000-671 Évora, Portugal
ipr@di.uevora.pt

ABSTRACT

We present a Question-Answering (QA) system for Portuguese juridical documents.

The QA system was applied to the complete set of decisions from several Portuguese juridical institutions (Supreme Courts, High Court, Courts, and Attorney-General's Office) in a total of 180,000 documents.

1. INTRODUCTION

Question answering systems [5] are an important topic of research in the natural language processing field and the legal domain is an area where question answering systems could (and should) be applied, allowing citizens to have an easier access to legal information[4].

One way of improving the criminal investigation is to enable the investigators to find similar criminal processes helping them to learn from the mistakes that were committed in cases that have prescribed or where the prosecution has lost in court. In order to overcame the lake of a structured documents database with the criminal process's documentation, we propose the use of a question answering system with the following goals:

• Answering user questions, using the information contained in the criminal process's, about:

Places Ex: Where are travesty bars in Lisbon? Where can we buy drugs?

Dates Ex: When was Mr. X arrested? When was built X building?

Definitions Ex: What is a drug dealer? What is a travesty bar?

Specific Ex: How many times was Mr X accused? Who was arrested by dealing drugs in process X? What crimes committed Mr Y?

• Indicate a set of relevant process's (documents).

Some times the investigator is not interested in obtaining just an answer to a questions; he/she may want to find the knowledge source of the system for answering a question.

ICAIL '05 June 6-11, 2005, Bologne, Italy

2. SYSTEM OVERVIEW

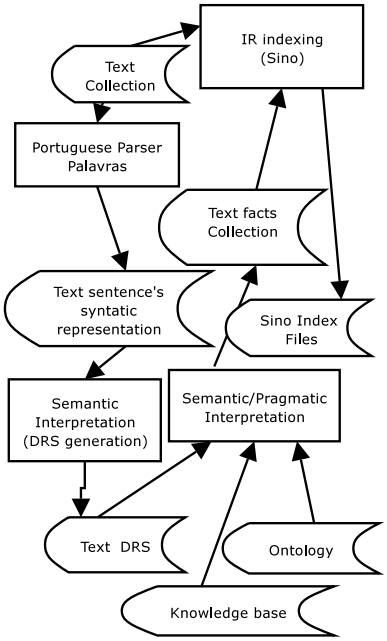

Figure 1: Texts preprocessing

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.

2.1 Processing the database documents

In this phase the main tasks are:

• **1** *Information retrieval indexing* – creates the files that index the full set of documents.

• **2** *Portuguese Parser* – each text of the collection is analyzed by the Portuguese parser PALAVRAS [2]. We've chosen to keep the first syntactic analysis for each sentence.

• **3** *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)[3].

• **4** *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 [6]. The obtained ontology was created in the OWL

(Ontology Web Language) format and in a logic programming framework, ISCO [1].

• **5** *Semantic/Pragmatic Interpretation* – gives rise to a set of knowledge bases, where each knowledge base has the facts conveyed by each text.

2.2 Query answering

Due to the lack of structure of the documents and due to the large amount of information contained in a large set of documents, our system has three main steps:

• **1st step** – each query is deeply analyzed: parsed and semantic pragmatic interpreted.

"Quem cometeu um homicidio por negligência por conduzir alcoolizado?/Who committed manslaughter (negligent murder) for drunk driving?" is transformed into:

```
drs([who-A-X-Y, undef-B-m-s,
     undef-C-f, undef-D-m-s],
    [committed(A,B), murder(B),
     rel(by,B,C), negligent(C),
     rel(for,A,D), drunk(D), drive(D)]
```

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:

```
drs([who-A-X-Y, def-B-m-s, def-C-m-s],
    [manslaughter(A,T), drive(A,T,_,_,C),
     alcool_degree(C),C>0.5]
```

This final query representation will be evaluated in each knowledge base selected by the information retrieval module.

• **2nd step** – documents selection, an information retrieval system (SINO) selects a set of potentially relevant documents. This will reduce the number of documents (kbs) for the next step.

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. This information retrieval task is performed using not only the words used in the question but also taking into account the query syntactic parser and its semantic/pragmatic interpretation.

• **3rd step** – Answer inference process, for each document selected, the semantic pragmatic representation of the query is evaluated in the document semantic/pragmatic representation.

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.

3. SYSTEM EVALUATION

Evaluation of the QA system was done by building a set of questions and, after this step, we experimentally validated the system's answers.

The system's answers always indicates: one or more natural languages terms and the document reference where the system finds out the answer as well a transcription of the sentence used to obtain the answer.

Testes were done using 200 questions and answer validation was done using the following classification:

• Correct answer and well supported. Indicating that the answer was correct and was supported in an adequate document. 25% of the questions obtained this classification

• Correct answer but not well supported. Indicating that the answer was correct but it was not supported in an adequate document. 2% of the answers obtained this classification

• Incorrect answer but the document and the sentence selected contained the answer. Indicating that the answer was not correct but the system was able to select an adequated document. 18% of the answers were classified this way.

• Incorrect answer and the sentence that was used by the system did not contain any answer to the question. 9% of the the answers were wrong.

• The system did not answer the question 46% of questions did not obtain any answer.

These results show an overall accuracy of around 25% correct answers. The analysis of the 46% of cases where the system could not obtain any answer showed that the main cause is due to lack of knowledge of the system: wrong syntactic analysis (of the question or of the document sentences that could contain the answer); lack of information on synonyms and incomplete ontology information; and finally the accuracy of the used information retrieval system. The main reason for not obtaining any answering (77% of the 46%) was the poorly performance of the used information retrieval system for selecting the potential set of documents that could contain the answer.

From the evaluation of the Question Answer System we can conclude the our system will not mislead the users, since the user can confirm the sources of knowledge of the system. The user just has to read the sentence used by the system to compute the answer in order to be able to classify the system's answer.

4. REFERENCES

[1] S. Abreu, P. Quaresma, L. Quintano, and I. Rodrigues. A dialogue manager for accessing databases. In *13th European-Japanese Conf on Information Modelling and Knowledge Bases*, pages 213–224, Kitakyushu, Japan, June 2003. Kyushu Inst of Tech.

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

[3] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Dordrecht: D. Reidel, 1993.

[4] M-F Moens. Interrogating legal documents: The future of legal information systems? In *Proc of the JURIX 2003 Work. on Question Answering for Interrogating Legal Documents*, pages 19–30. Utrecht University, 2003.

[5] P. Quaresma and I. Rodrigues. Using dialogues to access semantic knowledge in a web legal IR system. In M-F Moens, editor, *Procs. of the Work. on Question Answering for Interrogating Legal Documents of JURIX'03*, Utrecht, Netherlands, December 2003.

[6] J. Saias and P. Quaresma. Using nlp techniques to create legal ontologies in a logic programming based web information retrieval system. In *Work. on Legal Ontologies and Web based legal information management of the ICAIL03*, Edinburgh, Scotland, June 2003.