Using dialogues to access semantic knowledge in a web IR system

Paulo Quaresma and Irene Rodrigues

pq|ipr@di.uevora.pt,
Departamento de Informática,
Universidade de Évora, Portugal

Abstract. We present a dialogue system that enables the access in natural language to a web law information retrieval system. We use a semantic web language to model the document knowledge and to define an ontology representing the main classes of domain objects, their properties and their relations. Our system includes an ontology in DAML+OIL language describing the documents structure, a document database build with a subset of the Portuguese Attorney General's Office documents were each document includes a field with its semantic content described in DAML+OIL. The dialogue system interprets the user natural language sentences using the ontology, the documents semantic content and the inferred user attitudes.

1 Introduction

Our system was developed in order to supply some tools that enable citizens to access information in documents using natural language sentences queries.

The main problem in building such a system is how to obtain the knowledge necessary to perform the different analysis stages of natural language sentences.

We use semantic web language to model the document knowledge and to define an ontology representing the main classes of domain objects, their properties and their relations.

Using this ontology the documents are analyzed and their semantic content is represented. At the moment documents semantic content are only partially represent. We still need to build a more complete ontology and more powerful tools to extract the documents semantic contents.

We are using the DAML+OIL (Darpa Agent Markup Language - [9]) language which is defined using the RDF (Resource Description Framework - [5, 1]) language and it has a XML version.

Using DAML+OIL it is possible to represent the documents structure and some of its semantic content. Moreover, the user natural language queries can be semantically analyzed accordingly to the predefined ontology.

With this approach, the dialogue system that we propose is able to supply adequate answers to the Portuguese Attorney General's Office documents database (PGR).

For instance, in the context of an user interrogation searching for information on state pensions for relevant services to the country, the following question could be posed:

Who has pensions for relevant services?

The user is expecting to have as an answer some characteristics of the individuals that have that kind of pensions. He does not intend to have a list of those individuals or the documents that refer to the act of attributing such pension.

In order to obtain the adequate answer our system must collect all individuals referred in the documents database that have those pension and then it must supply the common characteristics to the user in a dialogue.

The answer to the above question could be:

'Individuals that were agents of an action putting their lives at risk'

This information can be obtained by extracting what those individuals have in common or by reading the Portuguese law. The possibility of extracting what are the common characteristics of a set of objects is a powerful tool for presenting the answers to our users. This behaviour can be achieved by choosing an adequate ontology to represent the objects including events present at the documents.

2 The Dialogue System

2.1 Web semantics

We have selected a domain of the Portuguese Attorney General documents – pensions (granted or refused) – and, in this domain, we have selected smaller sub-domains, such as, pensions for firemen, and militaries.

As an example, the *Individual* class is presented below (only some of the class attributes are shown):

```
<daml:Class rdf:ID="Individual">
        <daml:label>Individual</daml:label>
</daml:Class>
<daml:DatatypeProperty rdf:ID="individualCode">
 <daml:domain rdf:resource="#Individual"/>
 <daml:type rdf:resource=
         "http://www.w3.org/2001/03/daml+oil#UniqueProperty"/>
 <daml:range rdf:resource="http://www.w3.org/2000/10/XMLSchema#integer"/>
```

The two steps (ontology + document semantic representation) are the basis of the proposed system and allow the implementation of other steps, such as, the semantic/pragmatic interpretation and the dialogue management.

2.2 Natural Language Dialogue System

The analyses of a natural language sentence is split in four subprocesses: Syntax, Semantics, Pragmatics, Dialogue manager.

Syntax Analysis: our syntactic interpreter was built using *Chart Parsers*[2]. This is one of many techniques to build syntactic interpreters.

As an example, the following sentence:

"Who has a pension for relevant services?"

Has the following structure:

```
phrase([np([det(who,_+_+_), n('individual',_+s+m)]), vp(v('have',3+p+_)),
           args_v([np([det(a,_+p+_), n('pension',_+s+_),
             pp(for,np([name('relevant services',_+s+m)]))])])).
```

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)[4].

For instance, the semantic representation of the sentence above is the following expression:

individual(A), pension(B), name(C,'relevant services'), rel(B,C), have(A,B).

and the following discourse referents list:

[ref(A,p+_+_,what),ref(B,s+_+_,undef),ref(C,p+_+_,undef)]

The Semantic/Pragmatic Interpretation 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" [3]. This process was described in detail in [8].

The result of interpreting the above sentence is the following expression:

pension_relevant_services(B,A,_,_,_), individual(A,_,_,_). Where:

- $B =_{\#} (1046..1049 : 1345 : 1456..1457)$ – B constrained to all pension for relevant services.
- $A =_{\#} (7001...7852)$ – A is constraint to individuals

The above LP expression contain the possible interpretations of the sentence in the context of our documents database.

The Dialogue Manager must recognize the speech act associated with the sentence (in this domain it can be an $inform$, a $request$, or a $askif$ speech act), to model the user attitudes (intentions and beliefs), and to represent and to make inferences over the dialogue domain.

In order to achieve this goal the system needs to model the speech acts, the user attitudes (intentions and beliefs) and the connection between attitudes and actions. This task is achieved through the use of logic programming framework rules (see [7, 6] for a more detailed description of these rules).

In this framework, after having accessed the textbase, the system may have a multiple solution and it may need to start a clarification sub-dialogue. In the clarification sub-dialogue, the systems asks the user to select one of the possible solutions. In order to collaborate with the user we have defined a cluster predicate

that tries to aggregate the solutions into coherent sets. The strategy behind this predicate is to aggregate the solutions accordingly with the range of property values of the selected objects. For instance, in the presented example the selected individuals might be clustered by their profession, or by their support documents, or by the events in which they are actors.

Considering the already presented question, the answer might be:
'Individuals that are firemen, and militaries'.
Or, using another property (event list):
'Individuals that were agents of an action putting their lives at risk'

3 Conclusions and Future Work

The dialogue system described in this paper is still a prototype but it will be made available to all users in the context of the Portuguese Attorney General's web information retrieval system (http://www.pgr.pt).

Clearly, and due to its complexity, many modules have aspects that may be improved:

- The coverage of the semantic analyzer (plurals, quantifiers, ...)
- The ontology coverage
- The semantic representation of the documents content
- The capability of the dialogue manager to take into account previous interactions and the user models

References

1. D. Brickley and R. Guha. *Resource Description Framework (RDF) - Schema Specification.* W3C, 1999.
2. Gerald Gazdar and Chris Mellish. *Natural Language Processing in PROLOG.* Addison-Wesley, 1989.
3. 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.
4. H. Kamp and U. Reyle. *From Discourse to Logic.* Kluwer, Dordrecht, 1993.
5. O. Lassila and R. Swick. *Resource Description Framework (RDF) - Model and Syntax Specification.* W3C, 1999.
6. P. Quaresma and J. G. Lopes. Unified logic programming approach to the abduction of plans and intentions in information-seeking dialogues. *Journal of Logic Programming*, 54, 1995.
7. Paulo Quaresma and Irene Rodrigues. Using logic programming to model multi-agent web legal systems – an application report. In *Proceedings of the ICAIL'01 - International Conference on Artificial Intelligence and Law*, St. Louis, USA, May 2001. ACM. 10 pages.
8. Luis Quintano, Irene Rodrigues, and Salvador Abreu. Relational information retrieval through natural lanaguage analysis. In *Proceedings of INAP'01*, Tokyo, Japan, October 2001. INAP.
9. www.daml.org. *DAML+OIL – DARPA Agent Markup Language*, 2000.