Semantic Enrichment of a Web Legal Information Retrieval System

José Saias Paulo Quaresma
Departamento de Informática,
Universidade de Évora,
7000 Évora, Portugal
{jsaias,pq}@di.uevora.pt

Abstract. Intelligent text information retrieval systems need the capability to deal with the semantics of the content of their text bases. In order to satisfy this requisite it is necessary to extract semantic information from the documents and to be able to make inferences about it.

A methodology to semi-automatically transform a traditional web IR system into a semantic aware one is proposed. The methodology is composed by three major steps: construction of an appropriate semantic ontology; text enrichment with semantic information; and construction of the inference engine. In order to create an adequate ontology, natural language processing techniques are applied, such as, partial parsers and lexical information (WordNet). Documents are enriched with semantic information using the output of the partial parsers and the obtained ontology. Finally, an inference engine based on a declarative programming language – Prolog – is used as the basis for the reasoning process.

An application of this methodology to the legal web information retrieval system of the Portuguese Attorney General's Office is described.

1 Introduction

Intelligent text information retrieval systems (TIR) need the capability to deal with the semantics of the content of their text bases. In fact, there is a need for a shift from word-based TIR systems to content-aware ones. This shift is the basis of the semantic-web languages, such as, RDF (Resource Description Framework - [6]), SHOE (Simple HTML Ontology Extensions - [5]), and DAML+OIL (Darpa Agent Markup Language - [10]). In order to satisfy this requisite it is necessary to extract semantic information from the documents, to represent it, and to be able to make inferences about it.

A methodology to semi-automatically transform a traditional web IR system into a semantic aware one is proposed. The methodology is composed by three major steps:

- Construction of an appropriate semantic ontology;
- Text enrichment with semantic information;
- Construction of an inference engine able to reason about the semantic information.

In order to create an adequate ontology, natural language processing techniques are applied, such as, partial parsers and lexical information (WordNet). The approach described in this paper has similarities with the work that is being done by many researchers in the

José Saias and Paulo Quaresma, 'Semantic Enrichment of a Web Legal Information Retrieval System' in T.J.M. Bench-Capon, A. Daskalopulu and R.G.F. Winkels (eds.), *Legal Knowledge and Information Systems. Jurix 2002: The Fifteenth Annual Conference*. Amsterdam: IOS Press, 2002, pp. 11-19.

domain of semi-automatic enrichment of electronic documents. For instance, there are similarities with the work of Woods [9] at SUN labs on automatic conceptual indexing. In fact, we have also developed an approach that tries to extract semi-automatically an ontology from a set of texts and we also use the notion of most specific subsumer (MSS) in our searches. In this way, our system is able to use the semantic relationships between words and concepts to establish connections between the queries and the documents content. As it is described in the next section, we used a syntactical analyzer for the Portuguese language [1], which enabled us to detect and to extract the set of verbs and noun forms. One of the main differences between our approach and Woods' approach is the fact that we tried to improve the resulting ontology merging it with an already existent ontology for the legal domain, developed previously by the Portuguese Attorney General's Office.

The ontology creation task has also similarities with the the work of Engers et al. [2] in the context of the IST programme E-POWER and E-COURT. In these projects, XML standards for the legal domain are being proposed. Our goal is quite different: we do not intend to propose specific standards for all the legal domain; we aim to define a methodology that allow us to semi-automatically transform a traditional IR system into a semantic aware one. In the context of our work we are potential users of the results of the E-POWER project, namely, of its proposed legal ontology.

After having defined the ontology, documents are enriched with semantic information using the output of the partial parsers. The approach proposed to extract and to represent semantic information can be related with the work described by Hausser [4], where he proposes the use of a database metaphor to represent natural language content. In our proposal we also intend to represent natural language semantics in a propositional database-like way.

Finally, an inference engine based on a declarative programming language – Prolog – is being used as the basis for the reasoning process. This step has also many links with the work that is being done by the W3C – World Wide Web Consortium (and others) in the context of the XML, XSLT, XPATH, and XQUERY languages[1]. In fact, in the context of another research project, we are developing a Prolog tool, which will be able to handle queries in the XQUERY format.

Section 2 describes the ontology creation; section 3 describes the text enrichment process; and section 4 describes the inference mechanisms. Finally, in section 5 some conclusions and future work are pointed out.

2 Ontology Creation

The first two major steps in the construction of an ontology are the definition of:

- the knowledge domain
- the semantic language used to represent the knowledge

In previous work Quaresma and Rodrigues [8, 7] have described the legal web information retrieval system of the Portuguese Attorney General's Office. This system is available in the web (http://www.pgr.pt) and it has around 7,000 documents and 10,000,000 words (in Portuguese). Documents have a specific structure and they are defined in an XML-compatible format. As a consequence of the availability of this system, the legal domain was chosen to be used in the ontology construction. However, the legal domain is a very general one and, in the scope of this work, a subset of the legal domain was selected to be represented. This subset was selected taking into account its relevance and the existence of documents about it.

[1] See http://www.w3.org

In order to represent the legal knowledge there is a need for a semantic language able to represent and to reason about ontologies. Moreover, the semantic language should also be web-compatible and, as a consequence, XML-based. DAML+OIL, Darpa Agent Markup Language - [10], was chosen to be the base representation language because it has all these features and it was already defined a standard version[2].

Having chosen the language and the domain, the next step was to create an appropriate ontology. The ontology should represent the concepts of the domain, and their characteristics and relations.

The ontology creation process was divided in two steps:

- Definition of structural objects
- Definition of content (semantic) objects

By structural objects, we mean objects that can be inferred from the structure of the documents; by content objects, we mean objects that can be inferred from the content (semantics) of the documents.

The structural objects intend to capture what Boer et al. [2] call the form of the legal document. The semantic objects intend to model the legal documents content.

2.1 Structural Objects

The first kind of objects, structural objects, was defined after analysing the structure of the Portuguese Attorney General's Office documents. Two classes were identified as fundamentals, having several fields (or attributes):

- Document
- Classification

The first class, *Document*, has a set of attributes and each instance will be associated to a specific document. As an example, the daml+oil definition of this class is presented[3]:

```
<daml:Class rdf:ID="Document">
<daml:label>Document</daml:label>
</daml:Class>

<daml:DatatypeProperty rdf:ID="numDoc">
<daml:domain rdf:resource="#Document"/>
<rdf:type rdf:resource="http://www.daml.org
         /2001/03/daml+oil#UniqueProperty"/>
<daml:range rdf:resource="http://www.w3.org
         /2000/10/XMLSchema#string"/>
</daml:DatatypeProperty>
```

With this daml+oil code the class *Document* was defined having as property the attribute document number *numDoc*.

The other structural class is the class *Classification*, which is used to represent the classifications, or subjects, that characterise the documents. These legal subjects are connected by an hierarchy of relations, such as: moreGeneralThan, moreSpecificThan, relatedSubject, equivalentSubject.

The daml+oil definition of this class is:

[2] We intend to use the new proposed ontology language, OWL – Web Ontology Language, as soon as the W3C clearly defines its standard.

[3] The complete set of attributes is not presented in this paper due to space constraints.

```
<daml:Class rdf:ID="Subject">
<daml:label>Subject</daml:label>
</daml:Class>

<daml:DatatypeProperty rdf:ID="name">
<daml:domain rdf:resource="#Subject"/>
<rdf:type rdf:resource="http://www.daml.org/
           2001/03/daml+oil#UniqueProperty"/>
<daml:range rdf:resource="http://www.w3.org/
          2000/10/XMLSchema#string"/>
</daml:DatatypeProperty>

<daml:ObjectProperty rdf:ID="moreGeneralThan">
<daml:domain rdf:resource="#Subject"/>
<daml:range rdf:resource="#Subject"/>
</daml:ObjectProperty>

<daml:ObjectProperty rdf:ID="moreSpecificThan">
<daml:inverseOf rdf:resource="#moreGeneralThan"/>
</daml:ObjectProperty>
```

The complete hierarchy of legal subjects was automatically built from an already existent version developed manually by the Portuguese Attorney General's Office. The complete hierarchy has around 6,000 concepts and 9,000 *moreSpecificThan* relations. Our partners from the Portuguese Attorney General's Office are still working in this ontology trying to improve it and to define new links.

An example of the daml+oil representation of a legal subject with some relations is:

```
<pgr:Subject rdf:ID="c7276">
  <pgr:code>7276</pgr:code>
  <pgr:name>Accident</pgr:name>
  <pgr:moreGeneralThan rdf:resource="#c1346"/>
  <pgr:moreGeneralThan rdf:resource="#c1348"/>
  <pgr:moreSpecificThan rdf:resource="#c7275"/>
</pgr:Subejct>
```

2.2 Semantic Objects

The definition of the second kind of objects, semantic objects, is a more complex task and it is still under development. We have identified several subtasks:

1. Identification of the most important verbs and nominal expressions;
2. Selection of a subset of the identified verbs and nominal expressions;
3. Characterisation of the selected verbs and nominal expressions;
4. Creation of the correspondent ontology

The first subtask, verb and nominal expressions identification, was done using the following approach:

- Text syntactical parsing. The documents were analysed by the parser developed by E. Bick in the domain of the VISL project (http://visl.hum.sdu.dk/visl [1]).

- Verb and nominal expressions extraction. Using the parser output, an analyser was developed in Prolog, which is able to extract the verbs and nominal expressions from the sentences.
- Verb and nominal expressions frequency. The verb and nominal expressions frequency for the complete set of documents was computed.

The second subtask, verb and nominal expressions selection, was done in a semi-manual way. First, the top verbs and nominal expressions were selected from the list of the identified verbs and nominal expressions. Then, the verbs and nominal expressions with correspondent concepts in the legal hierarchy were also selected. Finally, these lists was verified by legal experts. As an example, the first identified verbs were:

- ser – to be – *not selected*
- ter – to have – *not selected*
- ...
- referir – to refer – *selected*
- aprovar – to approve – *selected*
- ...

The third subtask, characterisation of the selected verbs and nominal expressions, was done using the following approach:

- For each verb occurrence, the subject and the direct object were extracted;
- For each nominal expression occurrence, the correspondent verb was extracted;
- Normalisation of the extracted verbs, subjects and direct objects. In this phase, we were able to identify new concepts and instances, such as, the agents that perform actions or the entities that are direct objects of these actions.

We tried to relate the new concepts using the results of the WordNet project, which semantically relates lexical entities. However, the Portuguese version of the WordNet has few entries and it was not possible to build many relations automatically. In the future, we intend to use the English version as the basis for the construction of concept relations.

The fourth subtask, was the creation of the ontology of the extracted concepts: verbs, subjects, and direct objects.

A subset of the daml+oil result code is:

```
<daml:Class rdf:ID="Action">
<daml:label>Action</daml:label>
</daml:Class>

<daml:Class rdf:ID="Entity">
<daml:label>Entity</daml:label>
</daml:Class>

<daml:ObjectProperty rdf:ID="subject">
<daml:domain rdf:resource="#Action"/>
<daml:range rdf:resource="#Entity"/>
</daml:ObjectProperty>
```

```
<daml:ObjectProperty rdf:ID="object">
<daml:domain rdf:resource="#Action"/>
<daml:range rdf:resource="#Entity"/>
</daml:ObjectProperty>

<pgr:Action rdf:ID="a1">
  <pgr:code>1</pgr:code>
  <pgr:name>to approve</pgr:name>
</pgr:Action>

<pgr:Entity rdf:ID="e142">
  <pgr:name>Portuguese Attorney General</pgr:name>
</pgr:Entity>

<pgr:Entity rdf:ID="e21">
  <pgr:name>Law</pgr:name>
</pgr:Entity>
```

This code defines the classes *Action* and *Entity* and relates them through the *subject* and *object* property. Moreover, it gives an example of the action *to approve* and entities *Portuguese Attorney General* and *Law*.

3 Semantic Text Enrichment

The second step in the proposed methodology is to transform the original documents into semantic web ones, or to enrich them with semantic information.

This task was divided in two subtasks:

- Structural information
- Semantical concepts

The first subtask was done through the use of a Java parser, which automatically processes the documents, detects the structural information (which was already XML-tagged) and inserts the correspondent daml+oil code. In order to create the correct links to the legal subjects this subtask needs to have information about the ontology proposed in the previous section. For instance, each document has instances of the adequate legal subjects of the ontology.

The second subtask, semantical concepts, used as input the parsed documents and the daml+oil ontology. For each verb in the ontology and in a specific document, an instance of the correspondent action with its subject and direct object was created. For example, the action *to approve* and the entities *Portuguese Attorney General* and *law* are related by the following links:

```
<pgr:Action rdf:ID="a1">
  <pgr:subject rdf:resource="#e142"/>
  <pgr:object rdf:resource="#e21"/>
</pgr:Action>
```

This link means that there is an instance of the action *to approve*, which has the Portuguese Attorney General as subject and the *law* as direct object.

All the generated daml+oil code was validated using the available daml+oil validator: ”http://www.daml.org/validator/”.

4 Semantic Inference Engine

The next step in the proposed methodology is the development of an inference engine able to handle questions about the semantic representation of the documents.

As final goal, we intended to handle the following kind of questions:

- Documents where property P is V
- Documents about the concept C
- Documents where action A is performed
- Documents where action A is performed having subject S
- Documents where S is the subject of an action
- ...

Note that the inference engine needs to be able to deal with the ontology relations. For instance, the question ”documents about concept C” means ”documents about concept C or any of its more specific concepts” and the question ”documents where action A is performed having subject S” means ”documents where action A (or any of its sub-classes) is performed having subject S (or any of its sub-classes)”.

As a consequence, the inference engine needs the capability to represent knowledge, namely ontologies, and to reason about the represented knowledge. As it was already stated, the final goal of this step is to develop and to use a XQUERY-compatible inference engine. At this project phase, we are using Prolog as the query language. but we intend to use in the future a full compatible XQUERY interpreter.

In order to be able to reason about the daml+oil concepts, the semantics of the daml+oil language [3] should be represented by Prolog rules. At present, we do not have the full daml+oil semantics represented by Prolog rules. In fact, we have only the rules needed for the subset of daml+oil that is used by our generators: classes, subclasses, and properties.

There are two Java translators:

- Ontology translator – daml+oil ontology → Prolog
- Document translator – daml+oil document instances → Prolog

The first translator, receives as input the daml+oil ontology created in section 2 and creates the correspondent Prolog facts and rules needed to model the daml+oil semantics. As an example, the daml+oil code for class *Document* presented in section 2 is represented by the following Prolog code[4]:

```
class(document, 'Document').

property(numDoc).
domain(numDoc, document).
range(numDoc, string).
type(numDoc, unique).
```

[4] Due to its complexity, the presented code is a simplified version of the actual Prolog code.

The second translator, document translator, receives as input the daml+oil documents and creates the correspondent Prolog facts. Basically, each daml+oil instance will have a correspondent Prolog fact. For example, the daml+oil code for the *to approve* action presented in section 3 is represented by the following Prolog facts:

```
action(a1).
property(a1, subject, e142).
property(a1, object, e21).
```

After the translation of the ontology and the documents into Prolog facts and rules it is possible to handle the questions presented earlier in this section:

- Documents where property P is V

  document(X), property(X, P, V).

- Documents about the concept C

  document(X), concept(Y), name(Y, C), property(X, concept, Y).

- Documents where action A is performed

  document(X), action(Y), name(Y, A), property(X, action, Y).

- Documents where action A is performed having subject S

  document(X), action(Y), name(Y, A), name(Z, S), property(X, action, Y), property(Y, subject, Z).

- Documents where S is the subject of an action

  document(X), name(Z, S), property(X, action, Y), property(Y, subject, Z).

As it can be seen, using the proposed methodology and the Prolog inference engine with its variable unification and backtracking mechanisms it is possible to answer queries about the semantic content of the text bases.

5 Conclusions and Future Work

A methodology to transform a classical text information retrieval system into a semantic web one is presented. This methodology allows us to semi-automatically create an ontology for the knowledge domain and to enrich the documents with the extracted semantic information. Moreover, a Prolog based inference engine was developed able to represent and to infer about the semantic knowledge.

However, this is an ongoing project and there are many areas where work needs to be done:

- Semantic concepts. The semantic concepts were automatically extracted from the result of a syntactical document parsing. However, the parsing process has errors and, as a consequence, there are errors in the semantic extraction.

- Normalisation of concepts. Some work was done trying to normalise the concepts used as subject or direct object of verbs, but this process did not eliminate all the duplicates and incorrections. More work needs to be done in this area.

- WordNet. The identified concepts were related through the access to WordNet, a lexical ontology. However, the Portuguese version of WordNet has many limitations. We intend to use the English version in the future (but we'll need to automatically translate our concepts and the results of WordNet).
- Daml+oil (or OWL) semantics. A full model of the daml+oil (or OWL) semantics needs to be implemented in Prolog.
- XQuery compatibility. The Prolog inference engine should be able to handle the XQuery language.
- Natural Language Interface. A natural language interface between the users and the Prolog inference engine is also going to be developed in the context of this project.
- Integration and evaluation. The proposed methodology needs to be integrated into a web information retrieval system and to be evaluated by the users.

Acknowledgements

We would like to thank the JURIX referees for their helpful comments on the first version of this paper.

References

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

[2] A. Boer, R. Hoekstra, R. Winkels, T. van Engers, and F. Willaert. Proposal for a dutch legal xml standard. In *EGOV2002 – Proceedings of the First International Conference on Electronuc Government*, 2002.

[3] Frank Harmelen, Peter Patel-Schneider, and Ian Horrocks. A model-theoretic semantics for daml+oil. Technical report, www.daml.org, 2001.

[4] Roland Hausser. Database semantics for natural language. *Artificial Intelligence*, 130:27–74, 2001.

[5] Jeff Heflin. *Towards the semantic web: Knowledge Representation in a Dynamic Distributed Environment.* PhD thesis, University of Maryland, College Park, USA, 2001.

[6] O. Lassila and R. Swick. *Resource Description Framework (RDF) - Model and Syntax Specification.* W3C, 1999.

[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.

[8] Paulo Quaresma and Irene Pimenta Rodrigues. Pgr: Portuguese atorney general's office decisions on the web. In Osamu Yoshie, editor, *Proceedings of the 14th International Conference on Applications of Prolog*, University of Tokyo, Tokyo, Japan, October 2001. REN Associates, Inc. ISSN 1345-0980. To be published by Springer Verlag's LNAI.

[9] W. Woods. Conceptual indexing: A better way to organize knowledge. Technical report, Sun Microsystems Laboratories, Mountain View, CA, 1997. Technical Report SMLI TR-97-61. http://www.sun.com/techrep/1997/abstract-61.html.

[10] www.daml.org. *DAML+OIL – DARPA Agent Markup Language*, 2000.