Legal Information Extraction ←
Machine Learning Algorithms + Linguistic Information

Paulo Quaresma

Computer Science Department & CENTRIA – AI Centre
School of Sciences and Technology,
University of Évora, Portugal
pq@di.uevora.pt

Abstract
In order to automatically extract information from legal texts we propose the use of a mixed approach, using linguistic information and machine learning techniques. In the proposed architecture, lexical, syntactical, and semantical information is used as input for specialized machine learning algorithms, such as, support vector machines. This approach was applied to collections of legal documents and the preliminary results were quite promising.

Keywords: Information Extraction, Semantic Analysis, Machine Learning Algorithms

1. Introduction
Information extraction from text documents is an important and open problem. Although this is a general domain problem, it has a special relevance in the legal domain. For instance, it is very important to be able to automatically extract information from documents describing legal cases and to be able to answer queries and to find similar cases. Much research work on this topic has been done in the last years, as it is described, for instance, in Stranieri and Zeleznikow’s book “Knowledge Discovery from Legal Databases” (Stranieri and Zeleznikow, 2005). Typical approaches vary from machine learning techniques, applied to the text mining task, to the use of natural language processing tools.

2. Proposal
We claim that a mixed approach, using deep linguistic information and machine learning techniques, is the best approach to handle this problem and to obtain good results. By “deep linguistic information” we mean lexical, syntactical and semantical information, linked with an ontology representing the knowledge of the domain.

The overall idea is to use natural language processing tools to analyse the legal texts and to obtain:

- Lexicon with part-of-speech (POS) tags;
- Syntactical parse trees;
- Partial semantical representation.

This linguistic information can be used as input features for specialized machine learning algorithms, which will be responsible for high-level information tagging and extraction. As machine learning techniques we propose the use of kernel-based ones, such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), which are able to handle complex structured data as input.

The extracted information – mainly legal concepts and named entities – can be used to populate domain ontologies, allowing the enrichment of documents and the creation of high-level legal information retrieval systems. These legal information systems are “semantic-aware” ones and they are able to answer queries about concepts, entities and events.

3. Example
As an example, suppose the legal text has the following (simple) sentence:

- The judge decided in favor of the plaintiff.

As natural language processing tools we have used NLTK – Natural Language Toolkit – from the University of Pennsylvania, “C&C” for the syntactical parser and “Boxer” for the syntactic to semantic representation (Curran et al., 2007; Bos, 2008).

Applying these tools to the presented example, we’ll obtain the following parse tree:

<table>
<thead>
<tr>
<th>The</th>
<th>judge</th>
<th>decided</th>
<th>in</th>
<th>favor</th>
<th>of</th>
<th>the</th>
<th>plaintiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>NN</td>
<td>VBD</td>
<td>IN</td>
<td>NNP</td>
<td>IN</td>
<td>DT</td>
<td>NN</td>
</tr>
<tr>
<td>DT</td>
<td>NN</td>
<td>VBD</td>
<td>IN</td>
<td>NNP</td>
<td>IN</td>
<td>DT</td>
<td>NN</td>
</tr>
</tbody>
</table>

Figure 1: Parse tree

In this tree it is possible to identify the noun phrase The judge, the main verb decided and the prepositional phrase in favor of the plaintiff. Applying the “Boxer” tool to the output of the C&C parser, we’ll obtain the discourse structure represented in figure 2. In this structure we have:

- two entities x0 and x1, which represent the judge and the plaintiff;
- the event x2, which represents the main action to decide and has an agent x0, the judge, and an object x3;

1 http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Demo
to specific problems reported in the cited paper). However, this work represented a first approach and our proposal was not fully applied, as we didn’t use semantic information.

In another more recent work (Gaspar et al., 2011) we’ve applied the complete “deep” linguistic analysis to a collection of texts from the Reuters dataset, obtaining a partial semantic representation of the sentences – as discourse representation structures (DRSs). These structures were represented by direct graphs. We have also proposed a graph kernel function able to calculate the similarity of the semantic structures and we used Support Vector Machines to classify texts. The results were promising with an accuracy higher than 50%. As referred in the previous section, as natural language processing tools we used the "C&C" syntactical parser and "Boxer" to obtain the semantic representation (Curran et al., 2007; Bos, 2008).

5. Conclusions

We proposed to use deep linguistic information and machine learning techniques to the legal information extraction task. The results obtained in preliminary results were quite promising. However, it is necessary to perform more experiences with bigger legal text collections.

6. References


