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[1], we'll obtain the following parse tree:

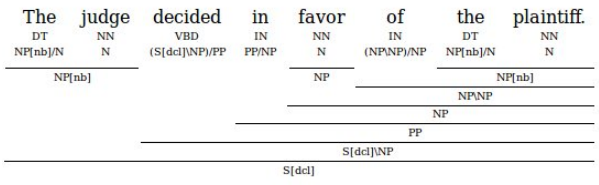

Figure 1: Parse tree

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

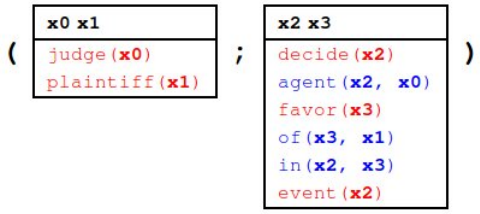

Figure 2: Discourse representation structure

- an object of the decision – *x3* –, which is related with *x2* and *x1* through the relations *in* and *of*.

As it was shown, the obtained output allows the automatic identification of entities and actions and the inference of their relations.
This information can be added as input features to machine learning algorithms aiming to improve the results of the named entity recognition task (NER) – identification of persons, organizations, and places.
Moreover, being a logically-based formalism, this representation allows an "easier" implementation of knowledge-based systems. In the presented example, suppose we have a legal ontology with the concept/class *person* with subclasses *judge* and *plaintiff*; it is possible to automatically populate these subclasses with instances *x0* and *x1*. The same approach can be applied to automatically populate instances of *events* and to create links between the created instances (as described in (Saias and Quaresma, 2005)). The obtained ontology (classes, instances, and relations) can be represented in a web language, such as OWL – Ontology Web Language, allowing its access through specialized semantic web search engines or using a query language, such as SPARQL (SPARQL Protocol And RDF Query Language).

4. Experiences

The use of deep linguistic information to automatically create and to populate legal ontologies was proposed and described in (Saias and Quaresma, 2005). This approach was completely based on symbolic natural language processing tools and it was applied to a collection of documents from the Portuguese Attorney General's Office. The limitations of the existent NLP tools (parsers, semantic analyzers) were one of the major reasons we've extended this approach to also use machine learning techniques.
In another previous work (GonÃ§alves and Quaresma, 2010) we have partially applied this methodology to a corpus of legal documents from the EUR-Lex site[2] within the "International Agreements" sections and belonging to the "External Relations" subject. The obtained results were very promising and, for the main concepts identification task, we obtained values higher than 95% for the precision and 90% for the recall. The identification of named entities (NER) showed also good results, varying from error rates of 0.1% for dates to around 15% in the identification of places and to a high value of 65% for organizations (due 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).

[2] http://eur-lex.europa.eu/en/index.htm

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

Johan Bos. 2008. Wide-coverage semantic analysis with boxer. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Research in Computational Semantics, pages 277–286. College Publications.

C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June. Association for Computational Linguistics.

Miguel Gaspar, Teresa Gonçalves, and Paulo Quaresma. 2011. Text classification using semantic information and graph kernels. In *EPIA-11, 15th Portuguese Conference on Artificial Intelligence*, pages 790–802, Lisbon, PT, October. ISBN: 978-989-95618-4-7.

Teresa Gonçalves and Paulo Quaresma. 2010. Using linguistic information and machine learning techniques to identify entities from juridical documents. In E. Francesconi, E. Montemagni, W. Peters, and D. Tiscornia, editors, *Semantic Processing of Legal Texts*, Lecture Notes in Artificial Intelligence 6036, pages 44–59. Springer.

Jose Saias and Paulo Quaresma. 2005. A methodology to create legal ontologies in a logic programming information retrieval system. In R. Benjamins, P. Casanovas, A. Gangemi, and B. Selic, editors, *Law and the Semantic Web*, Lecture Notes in Computer Science LNCS 3369, pages 185–200. Springer-Verlag.

A. Stranieri and J. Zeleznikow. 2005. *Knowledge Discovery from Legal Databases*. Law and Philisohy Library. Springer.