Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents

Paulo Quaresma and Teresa Gonçalves

Departamento de Informática, Universidade de Évora
7000-671 Évora, Portugal
{pq,tcg}@di.uevora.pt

Abstract. Information extraction from legal documents is an important and open problem. A mixed approach, using linguistic information and machine learning techniques, is described in this paper. In this approach, top-level legal concepts are identified and used for document classification using Support Vector Machines. Named entities, such as, locations, organizations, dates, and document references, are identified using semantic information from the output of a natural language parser. This information, legal concepts and named entities, may be used to populate a simple ontology, allowing the enrichment of documents and the creation of high-level legal information retrieval systems.

The proposed methodology was applied to a corpus of legal documents - from the EUR-Lex site – and it was evaluated. The obtained results were quite good and indicate this may be a promising approach to the legal information extraction problem.

Keywords: Named Entity Recognition, Natural Language Processing, Machine Learning.

1 Introduction

Information extraction from text documents is an important and quite open problem, which is increasing its relevance with the exponential growth of the "web". Every day new documents are made available online and there is a need to automatically identify and extract their relevant information.

Although this is a general domain problem, it has a special relevance in the legal domain. For instance, it is crucial to be able to automatically extract information from documents describing legal cases and to be able to answer queries and to find similar cases.

Many researchers have been working in this domain in the last years, and a good overview is done in Stranieri and Zeleznikow's book "Knowledge Discovery from Legal Databases" [1]. Proposed approaches vary from machine learning techniques, applied to the text mining task, to the use of natural language processing tools.

E. Francesconi et al. (Eds.): Semantic Processing of Legal Texts, LNAI 6036, pp. 44–59, 2010.

We propose a mixed approach, using linguistic information and machine learning techniques. In this approach, top-level legal concepts are identified and used for document classification using a well known machine learning technique – Support Vector Machines. On the other hand, named entities, such as, locations, organizations, dates, and document references, are identified using semantic information from the output of a natural language parser. The extracted information – legal concepts and named entities – may be used to populate a simple ontology, allowing the enrichment of documents and the creation of high-level legal information retrieval systems. These legal information systems will have the capacity to retrieve legal documents based on the concepts they convey or the entities referred in the texts.

The proposed methodology was applied to a corpus of legal documents from the EUR-Lex site[1] within the "International Agreements" sections and belonging to the "External Relations" subject. The obtained results were quite good and they indicate this may be a promising approach to the legal information extraction problem.

The paper is organised as follows: section 2 describes the main concepts and tools used in our approach – SVM for text classification and a syntactic/semantic parser for named entities recognition – and the document collection used to evaluate the proposal; section 3 describes the experimental setup for the identification of legal concepts task and evaluates the obtained results; section 4 describes the named entity recognition task and its results; section 5 briefly describes some related work; and, finally, section 6 presents some conclusions and points out possible future work.

2 Concepts and Tools

This section introduces the concepts and tools employed in this work: the machine learning text classification approach used to automatically identify legal concepts and the appliance of linguistic information for named entity recognition. It concludes by presenting the exploited juridic dataset.

2.1 Text Classification

The learning problem can be described as finding a general rule that explains data, given a sample of limited size. In supervised learning, we have a sample of input-output pairs (the *training sample*) and the task is to find a deterministic function that maps any input to an output such that the disagreement with future input-output observations is minimised. If the output space has no structure except whether two elements are equal or not, we have a *classification* task. Each element of the output space is called a *class*. The supervised classification task of natural language texts is known as *text classification*.

In text classification, documents must be pre-processed to obtain a more structured representation to be fed to the learning algorithm. The most common

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

approach is to use a bag-of-words representation, where each document is represented by the words it contains, with their order and punctuation being ignored. Normally, words are weighted by some measure of word's frequency in the document and, possibly, the corpus. Figure 1 shows the bag-of-words representation for the sentence "The provisions of the Agreement shall be applied to goods exported from South Africa to one of the new Member States.".

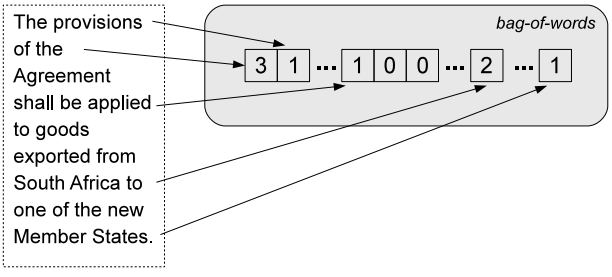

Fig. 1. Bag-of-words representation

In most cases, a subset of words (stop-words) is not considered, because their role is related to the structural organisation of the sentences, and does not have discriminating power over different classes. Some work reduces semantically related terms to the same root applying a lemmatiser.

Research interest in this field has been growing in the last years. Several machine learning algorithms were applied, such as decision trees [2], linear discriminant analisys and logistic regression [3], the naïve Bayes algorithm [4] and Support Vector Machines (SVM)[5].

[6] says that using SVMs to learn text classifiers is the first approach that is computationally efficient and performs well and robustly in practice. There is also a justified learning theory that describes its mechanics with respect to text classification.

Support Vector Machines. Support Vector Machines, a learning algorithm introduced by Vapnik and coworkers [7], was motivated by theoretical results from the statistical learning theory. It joins a kernel technique with the structural risk minimisation framework.

Kernel techniques comprise two parts: a module that performs a mapping from the original data space into a suitable feature space and a learning algorithm designed to discover linear patterns in the (new) feature space. These stages are illustrated in Figure 2.

The *kernel function*, that implicitly performs the mapping, depends on the specific data type and domain knowledge of the particular data source.

The *learning algorithm* is general purpose and robust. It's also efficient, since the amount of computational resources required is polynomial with the size and number of data items, even when the dimension of the embedding space (the feature space) grows exponentially [8].

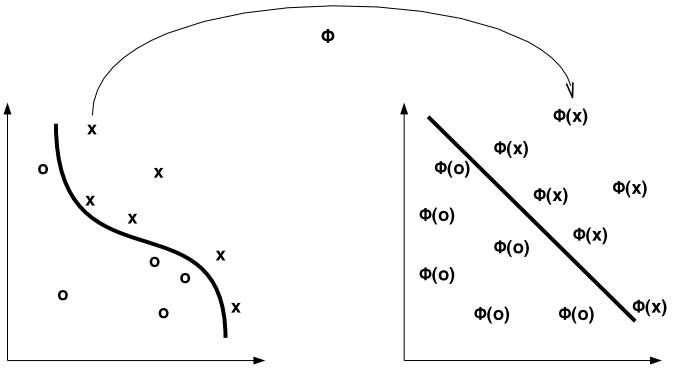

Fig. 2. Kernel function: data's nonlinear pattern transformed into linear feature space

Four key aspects of the approach can be highlighted as follows:

- Data items are embedded into a vector space called the feature space.
- Linear relations are discovered among the images of the data items in the feature space.
- The algorithm is implemented in a way that the coordinates of the embedded points are not needed; only their pairwise inner products.
- The pairwise inner products can be computed efficiently directly from the original data using the kernel function.

The *structural risk minimisation* (SRM) framework creates a model with a minimised VC (Vapnik-Chervonenkis) dimension. This developed theory [9] shows that when the VC dimension of a model is low, the expected probability of error is low as well, which means good performance on unseen data (good generalisation).

In geometric terms, it can be seen as a search to find, between all decision surfaces (the $\mathcal{T}$-dimension surfaces that separate positive from negative examples) the one with maximum margin, that is, the one having a separating property that is invariant to the most wide translation of the surface. This property can be enlighten by Figure 3 that shows a 2-dimensional problem.

SVM can also be derived in the framework of the regularisation theory instead of the SRM one. The idea of regularisation, introduced by [10] for solving inverse problems, is a technique to restrict the (commonly) large original space of solutions into compact subsets.

Classification Software. SVM^{light} [11] is a Vapnik's Support Vector Machine [12] implementation in C[2]. It is a fast optimization algorithm [6] that has the following features:

[2] Available at `http://svmlight.joachims.org/`

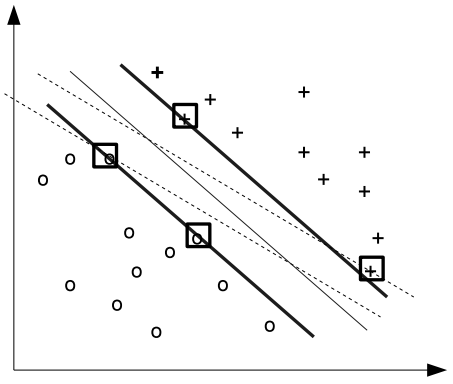

Fig. 3. Maximum margin: the induction of vector support classifiers

- solves classification, regression and ranking problems [13]
- handles many thousands of support vectors
- handles several hundred-thousands of training examples
- supports standard kernel functions and lets the user define your own

SVM^{light} can also train SVMs with cost models [14] and provides methods for assessing the generalization performance efficiently, the XiAlpha-estimates for error rate and precision/recall [15,6].

This tool has been used on a large range of problems, including text classification [16,5], image recognition tasks, bioinformatics and medical applications. Many of these tasks have the property of sparse instance vectors and using a sparse vector representation, it leads to a very compact and efficient representation.

2.2 Named Entity Extraction

A named entity extractor locates in text the names of people, places, organizations, products, dates, dimensions and currency. This information is needed to complete the final step in formation extraction of populating the attributes of a template. It is also useful to locate sentences that contain particular entities to answer questions.

To address this task machine learning techniques such as decision trees [17], Hidden Markov Models [18] and rule based methods [19] have been applied. In this work, instead of using a statistical approach, we will use a linguistic one.

Linguistic Information. The written language has a specific structure and comprehends several information levels. The most simple ones are the morphological and syntactic ones.

Morphological information includes a word's stem and its morphological features, like grammatical class and inflectional information. While some natural language processing tasks use a word's stem, others use its lemma.

Most syntactic language representations are based on the context-free grammar (CFG) formalism introduced by [20] and, independently, by [21]: given a sentence, it generates the corresponding syntactic structure. It is usually represented by a tree structure, known as sentence's *parse tree*, that contains its constituent structure (such as noun and verb phrases) and the grammatical class of the words.

Syntactic Parser Tool. Documents' syntactic structure was obtained using the PALAVRAS [22] parser for the English language. This tool was developed in the context of the VISL project by the Institute of Language and Communication of the University of Southern Denmark[3].

Given a sentence, the output is a parse tree enriched with some semantic tags. This parser is robust enough to always give an output even for incomplete or incorrect sentences, which might be the case for the type of documents used in text classification, and has a comparatively low percentage of errors (less than 1% for word class and 3-4% for surface syntax) [23].

For example, the output generated for the sentence "The provisions of the Agreement shall be applied to goods exported from South Africa to one of the new Member States." is

```
STA:fcl
=SUBJ:np
==>N:art("the" S/P) The
==H:n("provision" <act> <sem-c> <ss> <nhead> <left> P NOM) provisions
==N<:pp
===H:prp("of" <np-close>) of
===P<:np
====>N:art("the" S/P) the
====H:n("agreement" <sem-c> <act-s> <ss> <ac-cat> <nhead> S NOM) Agreement
=P:vp
==VAUX:v-fin("shall" <aux> PR) shall
==VAUX:v-inf("be" <aux>) be
==MV:v-pcp2("apply" <mv> PAS) applied
=PIV:pp
==H:prp("to" <right>) to
==P<:np
===H:n("goods" <cc-h> <nhead> P NOM) goods
===N<:icl
====P:v-pcp2("export" <mv> <np-close> PAS) exported
====ADVL:par
=====CJT:pp
======H:prp("from" <cjt-head> <advl-close> <right>) from
======P<:n("South_Africa" <complex> <nhead> <Proper> <Lcountry> S NOM) South_Africa
====P<<:pp
=====H:prp("to") to
=====P<:adjp
======H:num("one" <card> S) one
======N<:pp
=======H:prp("of" <np-close>) of
=======P<:np
========>N:art("the" S/P) the
========>N:adj("new" POS) new
========H:n("member_States" <complex> <nhead> <Proper> <heur> S NOM) Member_States
.
```

[3] Available at `http://www.visl.sdu.dk/`

2.3 Dataset Description

We performed the experiments over an set of European Union law documents. These documents were obtained from the **EUR-Lex** site[4] within the "International Agreements" section, belonging to the "External Relations" subject matter.

From all available agreements we chose the ones that had their full text (not just the bibliographic notice) and obtained a set of 2714 agreements dating from 1953 to 2008. Since the agreements are available in several languages we collected them for two anglo-saxon languages (English and German) and for two romanic ones (Italian and Portuguese), obtaining four different corpora: `eurlex-EN`, `eurlex-DE`, `eurlex-IT` and `eurlex-PT`.

Table 1 presents, for each corpus, the total number and average per document of tokens (running words) and types (unique words).

Table 1. Total number and average per document of tokens and types for each corpus

	tokens		types	
corpus	total	per doc	total	per doc
`eurlex-EN`	10699234	3942	73091	570
`eurlex-DE`	10145702	3728	133191	688
`eurlex-IT`	10665455	3929	96029	636
`eurlex-PT`	9731861	3585	86086	567

Each `eurlex` document is classified according to several ontologies: one obtained using the "EUROVOC descriptor", other using the "Directory code" and another using the "Subject matter". In all available classifications each document can be assigned to several categories. This setting is known as "multi-label".

The identification of legal concepts was accomplished using the first level of the "Directory code" classification, considering only the categories with at least 50 documents. Table 2 shows each category (id and name) along with the number of documents assigned to it.

Table 2. Number of documents assigned to each category

id	*name*	*# of docs*
2	Customs Union and free movement of goods	209
3	Agriculture	390
4	Fisheries	361
7	Transport policy	81
11	External relations	2628
12	Energy	58
13	Industrial policy and internal market	55
15	Environment, consumers and health protection	138
16	Science, information, education and culture	99

[4] Available at `http://eur-lex.europa.eu/en/index.htm`

3 Legal Concepts Identification

This section introduces the experimental setup and presents and evaluates the results obtained for the legal concepts identification task.

3.1 Experimental Setup

The experiments were done using a bag-of-words representation of documents, the SVM algorithm was run using SVM^{light} with a linear kernel and other default parameters and the model was evaluated using a 10-fold stratified cross-validation procedure.

Document Representation. To represent each document we used the bag-of-words approach, mapping all numbers to the same token and using the tf-idf weighting function normalised to unit length. This well known measure weights word w_i in document d as

$$\text{tf-idf}(w_i, d) = tf(w_i, d) \ln \frac{N}{df(w_i)}$$

where $tf(w_i, d)$ is the w_i word frequency in document d, $df(w_i)$ is the number of documents where word w_i appears and N is the number of documents in the collection.

Stratified Cross-validation. The cross-validation (CV) is a model evaluation method where the original dataset is divided into k subsets (in this work, $k = 10$), each one with (approximately) the same distribution of examples between categories as the original dataset (stratified CV). Then, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set; a model is built from the training set and then applied to the test set. This procedure is repeated k times (one for each subset). Every data point gets to be in a test set exactly once, and gets to be in a training set $k - 1$ times. The variance of the resulting estimate is reduced as k is increased.

Performance Measures. To measure learner's performance we analysed precision, recall and the F_1 measures [24] of the positive class. These measures are obtained from the contingency table of the classification (prediction *vs.* manual classification). For each performance measure we calculated the micro- and macro-averaging values of the top ten categories.

Precision is the number of correctly classified documents (true positives) divided by the number of documents classified into the class (true positives plus false positives).

Recall is given by the number of correctly classified documents (true positive) divided by the number of documents belonging to the class (true positives plus false negatives).

F_1 is the weighted harmonic mean of precision and recall and belongs to a class of functions used in information retrieval, the F_β-*measure*. F_β can be written as follows

$$F_\beta(h) = \frac{(1+\beta^2)prec(h)rec(h)}{\beta^2 prec(h) + rec(h)}$$

Macro-averaging corresponds to the standard way of computing an average: the performance is computed separately for each category and the average is the arithmetic mean over the ten categories.

Micro-averaging does not average the resulting performance measure, but instead averages the contingency tables of the various categories. For each cell of the table, the arithmetic mean is computed and the performance is computed from this averaged contingency table.

All significance tests were done regarding a 95% confidence level.

3.2 Results

While Figure 4 shows the micro- and macro-average precision, recall and F_1 graphically, Table 3 shows those measures for each category. For each measure, micro- and macro-average boldface values have no significant difference between them and the best value obtained.

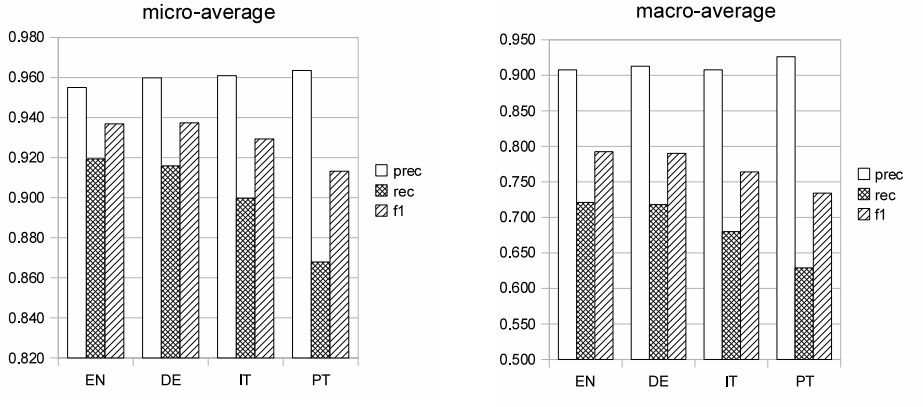

Fig. 4. Micro- and macro-average values

3.3 Evaluation

As can be seen in Figure 4, the precision values are good and the same for all studied languages (there's no significant difference between them): the micro-precision is above 0.95 while the macro one is above 0.90.

Having smaller values, the recall measure does not present the same behaviour: the best micro and macro-recall is for the English corpus, with .919 and .721 respectively, but while for the micro measure there is no significant difference

Table 3. Precision, recall and F_1 values for each category

	eurlex-EN			eurlex-DE			eurlex-IT			eurlex-PT		
id	prec	rec	F_1	prec	rec	F_1	prec	rec	F_1	prec	rec	F_1
2	.907	.651	.758	.952	.665	.783	.903	.579	.706	.929	.565	.702
3	.914	.818	.863	.926	.805	.861	.939	.705	.805	.942	.503	.656
4	.955	.934	.944	.965	.906	.934	.979	.914	.946	.971	.823	.891
7	.821	.568	.672	.846	.543	.662	.792	.519	.627	.813	.481	.605
11	.973	.998	.985	.973	.997	.985	.973	.998	.985	.973	.997	.985
12	.949	.638	.763	.872	.707	.781	.886	.672	.765	.921	.603	.729
13	.913	.382	.538	.895	.309	.459	.889	.291	.438	.944	.309	.466
15	.901	.725	.803	.918	.732	.815	.909	.725	.806	.902	.732	.808
16	.837	.778	.806	.868	.798	.832	.899	.717	.798	.941	.646	.766
micro	**.955**	**.919**	**.937**	**.960**	**.916**	**.937**	**.961**	**.900**	**.929**	**.964**	.868	.913
macro	**.908**	**.721**	**.792**	**.913**	**.718**	**.790**	**.908**	.680	.764	**.926**	.629	.734

for the German and Italian languages, for the macro one only the anglo-saxon languages present the best values.

Considering the individual category results, it is possible to conclude that the precision is always above recall for all languages and categories and as expected (since documents where retrieved having the "External relations" subject matter), the "External relations" category (id 11) have the best precision and recall with values almost equal to one in all languages. The "Fisheries" (id 4) also have very good values all above .9 (except the recall for the Portuguese corpus).

On the other way, there are some categories with small recall:

- while "Industrial policy and internal market" (id 13) has the worst ones, with values between .309 for the Portuguese corpus and .382 for the English one,
- "Transport policy" (id 7) has values between .481 for the Portuguese corpus and .568 for the English one and
- "Customs Union and free movement of goods" (id 2) and "Energy" (id 12) have values between .565 ("Customs" category for the Portuguese language) and .707 ("Energy" category for the German corpus).

Comparing results between languages, the English and German corpus present the best and very similar results, with the Portuguese one presenting the worst ones.

4 Named Entity Recognition

This section presents the experiments done for Named Entity Recognition. It begins by describing the experimental setup, then the results are presented and an evaluation is made.

4.1 Experimental Setup

The experiments were done using the `eurlex-EN` corpus (the collection for the English language). The following categories of Named Entities were studied:

- location names
- organization names
- dates
- references to documents and document articles

We did not try to extract personal names since after analysing the corpus we found almost no references to them.

For the extraction of location names and organization names we used the following subset of the semantic tags given by the parser PALAVRAS (see section 2.2):

- `<Lwater>`, `<Ltown>`, `<Lregion>` and `<Lcountry>` for location names
- `<HHorg>` and `<comp2>` for organization names

For the identification of dates we used a simple NLP tool, which received as input the sentences parse tree and performed a tree match procedure able to identify dates. References to other article and documents were also identified from the analysis of the parse trees.

After obtaining the candidate Named Entities, and since the corpus was not tagged, a manual evaluation was made for each category. For location names we made the analysis using the categorization given by PALAVRAS: "water" names (oceans, seas, rivers, etc. . .), towns, regions and countries.

4.2 Results

Table 4 shows for each kind of extracted named entities, the number of documents and for tokens (running words) and types (unique words) the total number and the minimum, maximum and average per document.

It is important to point out that we didn't obtain the number of unique references because we only identified and extracted the references inside the documents and we didn't try to consolidate the results. In order to be able to calculate this value we will need further text processing and it will be the focus of future work.

Table 4. Number of documents and for tokens (running words) and types (unique words) the total number and the minimum, maximum and average per document (for each kind of named entity)

		tokens				types			
category	docs	total	min	max	avg	total	min	max	avg
water	180	964	1	206	5.36	56	1	20	1.81
town	1820	11981	1	2001	6.58	307	1	54	2.32
region	1075	19438	1	456	18.08	220	1	46	2.77
country	2142	63979	1	621	29.87	521	1	97	4.72
organization	2281	56571	1	568	24.80	70	1	19	2.98
date	2714	19994	1	–	7.36	3521	1	–	1.29
reference	2714	76091	0	–	28.03	–	–	–	–

Table 5 presents the error percentage for each kind of named entity studied.

Table 5. Error percentage for each kind of named entity

category	error
water	12.5%
town	13.7%
region	18.2%
country	28.2%
organization	67.1%
date	0.1%
reference	65%

4.3 Evaluation

From table 4 we can state that these documents have a high number of references to other documents and articles (76091 references found and a 28% average per document). They also have high values of references to organizations and countries (56571 and 63979, respectively). These values are compatible with the type of analysed documents: legislation from the European Union. They also help to support our claim that this kind of information extraction is very important and it would allow the inference of important relations, such as, the chain of legislation references.

19994 date references were also identified, related to 3521 distinct events. This information can also be used as a basis for an analysis of relevant events in this legislation domain.

The performed evaluation focused on the precision of the information extraction modules and the results were shown in table 5. There are 3 classes of results:

- dates – The precision was quite good (error rate of 0.1%). This precision value was obtained because the legal documents have a quite standard way of presenting dates and a simple NLP tool was able to identify and extract the dates;
- location – Precision between 80 and 90%. These results depend heavily on the quality of the semantic tag classifier of the parser. We observed typical classes of errors and a simple upgrade of the parser's geographical information should significantly improve these results;
- organization and references – Precision around 35%. This quite low value has distinct explanations:
  - organization – the problem is caused by the semantic tag classifier of the parser. From a preliminary analysis it seems that all entities unknown to the system are classified as "organization". Only a change in the parser will allow an improvement of this result. Another approach might be to develop a special SVM classifier for this kind of entities.
  - reference – The high error rate value is explained by the complex syntactic structure used in the documents to make references to articles of other legislation. A deeper analysis of the syntactic sentence structure is needed to improve the quality of this sub-task.

5 Related Work

As referred in section 1 much work has been done in this domain in the last years. A good overview is done in the Stranieri and Zeleznikow's book "Knowledge Discovery from Legal Databases" [1]. In this book several approaches to the legal information extraction problem are described, varying from machine learning techniques to natural language processing methodologies. A more general but relevant reference in the information extraction domain is the "Information Extraction" paper of J. Cowie and W. Lehnert [25].

In the legal domain some of the related work is:

- [26] used decision trees to extract rules to estimate the number of days until the final case disposition;
- [27] developed rule based and neural networks legal systems;
- [28] used neural networks to model legal classifiers;
- [29,30] used SVM to classify juridical Portuguese documents;
- [31] proposed a framework for the automatic categorisation of case laws;
- [32,33] described the use of self-organising maps (SOM) to obtain clusters of legal documents in an information retrieval environment and explored the problem of text classification in the context of the European law;
- [34] described classification and clustering approaches to case-based criminal summaries;
- [35,36,37] described also related work using linear classifiers for documents;
- [38] integrated information extraction, information retrieval and machine learning techniques in order to design a case-based retrieval system able to find prior relevant cases. They used SVMs to rank prior case candidates.

6 Conclusions and Future Work

A proposal to identify and extract concepts and named entities in legal documents was presented and evaluated. The proposed methodology uses a SVM classifier to associate concepts to legal documents and a natural language parser to identify named entities, namely, locations, organizations, dates, and references to other articles and documents.

The concept classification task obtained an precision higher than 0.95 for the four languages selected in this experience (English, German, Italian, and Portuguese). Worst results were obtained for the romanic languages, which is compatible with previous research and is probably due to the use of more complex syntactic structures and richer morphology.

The named entities task obtained very good results for the identification of dates, an average result for locations (10-20% average error rate) and bad results for the identification of organizations and references to other articles and legislation. Extraction of locations can improve with the use of geographical databases and with the availability of this information to the parser – this will be the focus of future work. The identification of references to other articles and legislation needs a deeper analysis of the parse trees: from our error analysis we were able to

conclude that further work needs to be done in order to fully understand these syntactic structures.

Finally, we will improve our legal information retrieval system [39,40] to take into account the extracted information and to allow users to retrieve documents based on semantic information and not on surface-level words.

References

1. Stranieri, A., Zeleznikow, J.: Knowledge Discovery from Legal Databases. In: Law and Philisohy Library. Springer, Heidelberg (2005)
2. Tong, R., Appelbaum, L.: Machine learning for knowledge-based document routing. In: Harman (ed.) Proceedings of the 2nd Text Retrieval Conference (1994)
3. Schütze, H., Hull, D., Pedersen, J.: A comparison of classifiers and document representations for the routing problem. In: SIGIR 1995, 18th ACM International Conference on Research and Developement in Information Retrieval, Seattle, US, pp. 229–237 (1995)
4. Mladenić, D., Grobelnik, M.: Feature selection for unbalanced class distribution and naïve Bayes. In: ICML 1999, 16th International Conference on Machine Learning, pp. 258–267 (1999)
5. Joachims, T.: Transductive inference for text classification using support vector machines. In: ICML 1999, 16th International Conference on Machine Learning (1999)
6. Joachims, T.: Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers, Dordrecht (2002)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995)
8. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
9. Vapnik, V.: Statistical learning theory. Wiley, New York (1998)
10. Tikhonov, A., Arsenin, V.: Solution of Ill-Posed Problems. John Wiley and Sons, Washington (1977)
11. Joachims, T.: Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge (1999)
12. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
13. Joachims, T.: Optimizing search engines using clickthrough data. In: Schölkopf, B., Burges, C., Smola, A. (eds.) KDD 2002, 8th ACM Conference on Knowledge Discovery and Data Mining. ACM, New York (2002)
14. Morik, K., Brockhausen, P., Joachims, T.: Combining statistical learning with a knowledge-based approach – A case study in intensive care monitoring. In: ICML 1999, 16th International Conference on Machine Learning (1999)
15. Joachims, T.: Estimating the generalization performance of a SVM efficiently. In: Schölkopf, B., Burges, C., Smola, A. (eds.) ICML 2000, 17th International Conference on Machine Learning. MIT Press, Cambridge (2000)
16. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: ECML 1998, 10th European Conference on Machine Learning, Chemnitz, DE, pp. 137–142 (1998)
17. Baluja, S., Mittal, V., Sukthankar, R.: Applying Machine Learning for High Performance Named Entity Extraction. Computational Intelligence 16 (2000)

18. Miller, S.: Nymble: a high-performance learning name-finder. In: ALNP 1997, 5th Conference on Applied Natural Language Processing, pp. 194–201 (1997)
19. Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., Vilain, M.: MITRE: description of the Alembic system used for MUC-6. In: MUC6 1995, 6th Conference on Message Understanding, Morristown, NJ, USA, pp. 141–155. Association for Computational Linguistics (1995)
20. Chomsky, N.: Three models for the description of language. IRI Transactions on Information Theory 2, 113–124 (1956)
21. Backus, J.: The syntax and semantics of the proposed international algebraic of the Zurich ACM-GAMM Conference. In: Proceedings of the International Conference on Information Processing – IFIP Congress, UNESCO, Paris, pp. 125–132 (1959)
22. Bick, E.: The Parsing System PALAVRAS – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)
23. Bick, E.: A constraint grammar based question answering system for portuguese. In: Pires, F.M., Abreu, S.P. (eds.) EPIA 2003. LNCS (LNAI), vol. 2902, pp. 414–418. Springer, Heidelberg (2003)
24. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
25. Cowie, J., Lehnert, W.: Information extraction. Commun. ACM 39, 80–91 (1996)
26. Wilkins, D., Pillaipakkamnatt, K.: The effectiveness of machine laerning techniques for predicting time to case disposition. In: ICAIL 1997, 6th International Conference on Artificial Intelligence and Law, pp. 39–46. ACM, New York (1997)
27. Zeleznikow, J., Stranieri, A.: The split-up system: Integrating neural networks and rule based reasoning in the legal domain. In: ICAIL 1995, 5th International Conference on Artificial Intelligence and Law, pp. 194–195. ACM, New York (1995)
28. Borges, F., Borges, R., Bourcier, D.: Artificial neural networks and legal categorization. In: Proccedings of the 16th International Conference on Legal Knowledge Based Systems, pp. 11–20. IOS Press, Amsterdam (2003)
29. Gonçalves, T., Quaresma, P.: A preliminary approach to the multilabel classification problem of Portuguese juridical documents. In: Pires, F.M., Abreu, S.P. (eds.) EPIA 2003. LNCS (LNAI), vol. 2902, pp. 435–444. Springer, Heidelberg (2003)
30. Gonçalves, T., Quaresma, P.: Is linguistic information relevant for the classification of legal texts? In: Sartor, G. (ed.) Proceedings of the 10th International Conference on Artificial Intelligence and Law, Bologna, Italy, pp. 168–176. ACM, New York (2005)
31. Thompson, P.: Automatic categorization of case law. In: ICAIL 2001, 8th International Conference on Artificial Intelligence and Law, pp. 70–77 (2001)
32. Schweighofer, E., Merkl, D.: A learning technique for legal document analysis. In: ICAIL 1999, 7th International Conference on Artificial Intelligence and Law, pp. 156–163. ACM, New York (1999)
33. Schweighofer, E., Rauber, A., Dittenbach, M.: Automatic text representation, classification and labeling in european law. In: ICAIL 2001, 8th International Conference on Artificial Intelligence and Law, pp. 78–87 (2001)
34. Liu, C.L., Chang, C.T., Ho, J.H.: Classification and clustering for case-based criminal summary judgement. In: ICAIL 2003, 9th International Conference on Artificial Intelligence and Law, pp. 252–261 (2003)
35. Brüninghaus, S., Ashley, K.D.: Improving the representation of legal case texts with information extraction methods. In: ICAIL 2001: Proceedings of the 8th international conference on Artificial intelligence and law, pp. 42–51. ACM, New York (2001)

36. Brüninghaus, S., Ashley, K.: Finding factors: learning to classify case opinions under abstract fact categories. In: ICAIL 1997, 6th International Conference on Artificial Intelligence and Law, pp. 123–131. ACM, New York (1997)
37. Brüninghaus, S., Ashley, K.D.: Predicting outcomes of case-based legal arguments. In: ICAIL 2003, 9th International Conference on Artificial Intelligence and Law, pp. 233–242 (2003)
38. Al-Kofahi, A.K., Vachher, A., Jackson, P.: A machine learning approach to prior case retrieval. In: ICAIL 2001, 8th International Conference on Artificial Intelligence and Law, pp. 88–93 (2001)
39. Quaresma, P., Rodrigues, I.: A question-answering system for legal information retrieval. In: Moens, M.F., Spyns, P. (eds.) Proceedings of JURIX 2005: The 18th Annual Conference on Legal Knowledge and Information Systems, Frontiers in Artificial Intelligence and Applications, Brussels, Belgique, pp. 91–100. IOS Press, Amsterdam (2005)
40. Quaresma, P., Rodrigues, I.: A question-answering system for portuguese juridical documents. In: Sartor, G. (ed.) Proceedings of the 10th International Conference on Artificial Intelligence and Law, Bologna, Italy, pp. 256–257. ACM, New York (2005)