Using machine learning algorithms to identify named entities in legal documents: a preliminary approach

Prakash Poudyal , Luis Borrego and Paulo Quaresma

Departamento de Informática, ECT
Universidade de Évora, Portugal
{prakash,pq}@di.uevora.pt,luis_borrego@hotmail.com

Abstract. This paper deals with accuracy and performance of various machine learning algorithms in the recognition and extraction of different types of named entities such as date, organization, regulation laws and person. The experiment is based on 20 judicial decision documents from European Lex site. The obtained results were proposed for the selection of the best algorithm that selects appropriate maximum entities from the legal documents. To verify the performance of algorithm, obtained data from the tagging entities were compared with manual work as reference.

Keywords: Named entities recognition, Machine learning, Legal documents

1 Introduction

There are large scales of unstructured data stored in the web with numerous types of entities. To extract these entities from unstructured documents, information extraction algorithms are applied. However, it is quite difficult to know which algorithm is the best for particular types of entities. Therefore, it is of paramount importance to undertake experiments that could provide solutions to this problem.

The result from such study can be helpful to the lawyers, as a reference in cases when the retrieval of previous information is required. This easy way of accessing previous information may contribute towards the improvement of decision-making process.

Experiment is conducted in the default parameter of the algorithms in the minorthird [2] tool. There was no change in the parameters of the algorithms to obtain the result. Hence this experiment is given a tag of preliminary approach.

The paper is organized accordingly: in section 2 discusses related works regarding information extraction; in section 3 illustrates on concepts and tools that are

used for the experiment; in section 4 portrays on the experimental results and discussion; finally conclusion and future work is presented in the section 5.

2 Related works

Information extraction work is one of the important aspect of Machine learning: some previous works are discussion below.
The book "Knowledge Discovery from Legal Database" written by Stranieri and Zeleznikow[5] describes several approach of applying data-mining in law and also discusses trends in solving legal information extraction problem from machine learning techniques to natural language processing methodologies.

The article written by P. Quaresma and T. Goncalves [4] is the mixed approach, linguistic information and machine learning techniques to identify entities from judicial documents. Documents were available in four languages viz English, German, Italian and Portuguese. Top-level legal concepts are identified and used for document classification using support vector machine, where as named entities are identified using semantic information from output of a natural language parser.

Similarly, a book "Automatic Indexing and Abstracting of Document Texts" written by Marie-Francine Moens [3] emphasized in development of techniques for indexing and abstracting the text.

S. Baluja, V. O. Mittal and R.S.Hankar[1] presented a technique for named-entity extraction that automatically trained to recognize named-entities using statistical evidence from a training set.

3 Concepts and Tools

This section presents software used for the entity extraction, and description of dataset of judicial document.

3.1 Extraction Software

The machine-learning framework Minorthird[2]is open source software tool, which is collection of Java classes for storing text, annotating text, and learning extracting entities and categorizing text. All together 8 algorithms[1] were applied for the identification and extraction, which are listed below.

- Voted perceptron semi-Markov model (VPSMM)
- Voted perceptron conditional Markov model(VPCMM)
- Support vector machine conditional Markov model (SVMCMM)

[1] Note: Above listed algorithms are from javadoc of minorthird.

– Maximum entropy Markov model (MEMM)
– Conditional random fields (CRF)
– Semi-conditional random fields (SemiCRF)
– Voted perceptron hidden Markov model (VPHMM)
– Voted perceptron semi-Markov model 2(VPSMM2)

3.2 Dataset Description

Experiments were conducted in 20 judicial decision documents from the set of European Union law documents. These documents were obtained from EUR-Lex site[2]. The documents were available in several languages but for this experiment, English version was selected. Each document was splitted into 5 text files, which resulted in a total of 100 documents, because Minorthird suits in processing smaller text file rather than large. Entities that are extracted in this experiment listed below are:

– Name of person
– Name of organization
– Rules and Regulation Law
– Date that are available

These above entities are the most influential entities in judicial cases. The name of person, or the lawyer/criminal/judge in this case are important because they seem to appear more frequently for relevant searches, proving their influence in the related matter. The case for selection of the names of organizations, the rules and regulation laws and the dates of when the various activities occurred, a similar logic could be placed to emphasize their influence on the contextual topic. Hence these four entities have been prioritized and chosen over others. For extracting of above entities following subsets of semantic tags are given

Table 1. Entities with its semantic tags

Name of Entity	Semantic Tag
Date	<date></date>
Organization	<org></org>
Person	<person></person>
Regulation Laws	<rl></rl>

3.3 Experimental Setup

Experiments were conducted in minorthird and model was evaluated using a 10-fold stratified cross validation procedure.

[2] http://eur-lex.europa.eu/JURISIndex.do?ihmlang=en

Stratified Cross-validation: The cross-validation (CV) sometime called rotation estimation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set [6]. It is a model evaluation method where the original dataset is divided randomly partitioned into k subsets (in this experiment, k=10). Then, one of the k subsets is used as the test set and the other k-1 subsets are put together to form as a training set; a model is build from the training set and then applied to the test sets. This procedure is repeated k times (one for each subset). Every data get chance to be in a test set exactly once only, and gets to be in a training set k-1 times.

Performance Measures: To know the best algorithm we analyzed precision, recall and the F_1 measure of all entities. These three terms are described briefly.

Precision is defined as the number of relevant documents retrieved divided by the total number of documents retrieved of the positive class [7].

Recall is defined as the number of relevant documents retrieved divided by the total number of elements that actually belong to the positive class [7].

For example, there is total of 9 people in the corpus and system extract only 7 of them, out of which 4 contains the names of persons and 3 of dogs. In this case precision is 4/7 and recall will be 4/9

F-measure is the harmonic mean of precision and recall and belongs to a class of functions used in information retrieval. [7] F_β can be written as

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

4 Results and Discussion

The important two aspects are discussed here. The first is dealing with the identification of the best algorithm for each of the distinct entities and second is the comparison of the number of entities tagged by manual and machine.

Table 2 shows for each F-measure of Precision, Recall with F-measure. F-measure was considered to select the algorithm with highest value. In this case, Date has the highest value in Hidden Semi-Markov Models algorithm with f-measure value of 0.910 hence it is considered as the best algorithm but still support vector machine algorithm is competitive one with value 0.903. Similarly, Organization has highest value in Support vector machine with the value of 0.538. Similarly, Person has also highest value in Support vector machine with a 0.865. Regulation Law has highest value in Conditional Random Field with the value of 0.853.

Table 2. Precision, Recall and F measure values for each entity

	Date			Organization			Person			Regulation Laws		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
VPSMM2	.998	.096	.175	.000	.000	.000	.000	.000	.000	.000	.000	.000
VPCMM	.999	.225	.367	.446	.413	.429	.795	.339	.475	.927	.489	.640
CRF	.898	.820	.857	.659	.416	.510	.890	.840	.864	**.877**	**.831**	**.853**
SVMCMM	.898	.908	.903	**.646**	**.460**	**.538**	**.876**	**.854**	**.865**	.848	.848	.848
MEMM	.000	.000	.000	.000	.000	.000	.999	.037	.071	.000	.002	.005
SemiCRF	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
VPHMM	.000	.000	.000	.000	.000	.000	.967	.188	.315	.950	.495	.651
VPSMM	**.912**	**.908**	**.910**	.675	.441	.533	.803	.570	.667	.524	.055	.100

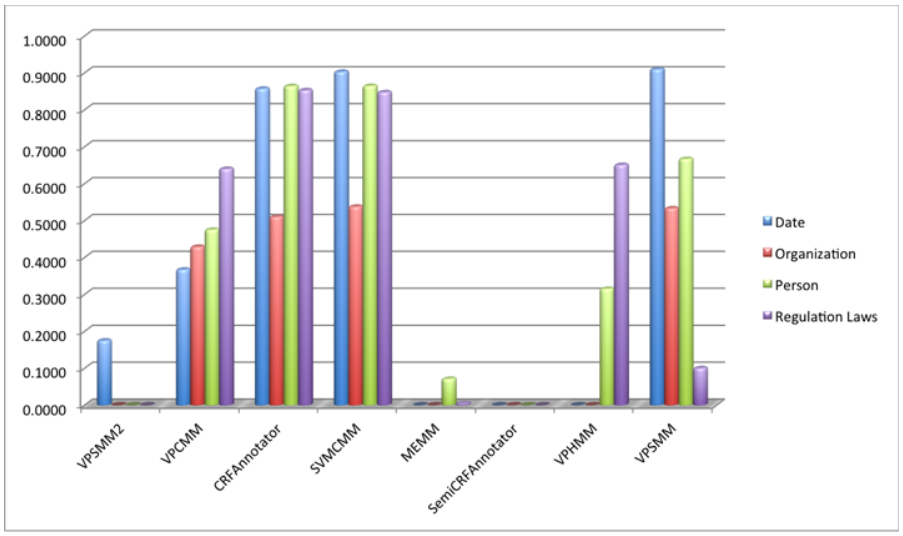

Fig. 1. bar chart of f-measure of each entities

4.1 Comparison between Manual Tagging and Machine Tagging

Manual tagging is all manual work that is conducted to tag these four entities. After conducting experiment in minorthird as explained in section 3.3; from f-measure of precision and recall, algorithm that is best to extract entities from judicial document was selected. For the verification of selecting, another setup of experiment conducted telling the respective algorithm to tag the entities in non tag judicial document, after all the result above in table 3 is more or less similar to the manual tagging number. So it can be believed by f-measure is quite promising to select the best algorithms.

Table 3. Compares manual tagging with system tagging

Entity	No. of Manual Tag	No. of Machine Tag	Algorithm	F-measure
Date	456	436	Hidden Semi Markov Model	0.910
Organization	411	395	Support Vector Machine	0.538
Person	534	531	Support Vector Machine	0.865
Regulation Laws	1388	1321	Conditional Random Field	0.853

5 Conclusion and Future work

In this paper, we have presented the results of a preliminary work aiming to automatically tag and extract information from juridical documents. The obtained results are quite promising and show that machine learning algorithms may be a good approach to deal with this problem. However, much work has to be done in order to improve the results and to be able to extract more information from the documents.

As future work, we also plan to create ontology able to represent legal knowledge and to automatically populate it with the information extracted from the legal documents.

References

1. Shumeet Baluja, Vibhu O. Mittal, and Rahul Sukthankar. Applying machine learning for high-performance named-entity extraction. *Computational Intelligence*, 16(4):586–595, 2000.
2. William W. Cohen. Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data, 2004.
3. Marie-Francine Marie-Francine Moens Moens. *Automatic Indexing and Abstracting of Document Texts*, volume 6 of *The Information Retrieval*. Springer, 2000.
4. Teresa Gonçalves Paulo Quaresma. Using linguistic information and machine learning techniques to identify entities from juridical documents. 6036:44–59, 2010.
5. Zeleznikow J. Stranieri A. *Knowledge Discovery from Legal Databases. In: Law and Philosophy*, volume 69 of *Law and Philosophy Library*. Springer, Heidelberg, 2005.
6. The free encyclopedia Wikipedia. Cross-validation (statistics) - wikipedia, the free encyclopedia, November 2011.
7. The free encyclopedia Wikipedia. Precision and recall - wikipedia, the free encyclopedia, November 2011.