Extracting information from court decisions

Gildas Tagny Ngompe[1,2], Sébastien Harispe[1], Jacky Montmain[1], Stéphane Mussard[2], Guillaume Zambrano[2]

[1] LGI2P, École des mines d'Alès, 69 Rue Georges Besse, Nîmes, France
[2] CHROME, Université de Nîmes, Rue du Dr Georges Salan, Nîmes, France

Abstract. We present our recent progress on the analysis of the French judicial corpus; we discuss in particular (i) results of our detailed study of HMM and CRF probabilistic models for detecting sections and entities in court decisions, and (ii) primilary work on the extraction of claim information through key-phase extraction and decision classification. We also discuss faced open challenges like dealing with multiple claims of similar or different types in the same decision, as well as lines of thought for solving these challenges based on discourse analysis.

Keywords: Natural Language Processing, probabilistic models, text classification, court decisions sectioning, entities and claims extraction

1 Introduction

A court decision is a document containing the description of a case, i.e. the decision of the judges as well as their motivations. In our aim to provide meaningful insights of court decisions corpora, we are designing automatic information extraction approaches enabling to extract relevant information for characterizing decisions and further analyse them. In particular, we have been working on (1) document sectioning considering three distinct parts (header, body, conclusion), (2) legal named entities detection (city, type of court, judges, parties, date, norm, lawyers, ...), and (3) claims extraction (i.e. for each claim made by parties, we aim at extracting its category, types of involved parties, requested quantum, result meaning, and granted quantum). More precisely, we have studied how well Hidden Markov Model (HMM) [1] and Conditional Random Fields (CRF) [2] can solve the two first tasks, and how results can be improved by taking into account additional labeled training data, as well as some particular design aspects like feature subset and segment representation selection. Ongoing work also focuses on experimenting a simple approach combining some weighting and classification methods to identify categories of claims and the meaning of the corresponding result; we also study how characteristic terms of a category can be used to locate the amount of money (quantum) requested and finally granted. This paper briefly provides details on our work and suggest some ways of improvement.

2 Sections and entities detection using HMM and CRF

HMM and CRF are well-known probabilistic models particularly adapted for extracting information from texts. In our work, we have studied their performance in detecting sections and entities in court decisions. To understand how HMM and CRF work, let's consider a text T as a sequence of observations or "*tokens*" $t_{1:n}$. Each t_i is a segment of text - in our case a word for entities and a line for sections. Considering a collection of labels L, labeling T consists of assigning the appropriate labels to each t_i. Our detection tasks are both segmentation tasks of text contents T, i.e. the aim is to split T into partitions such that the elements of a partition necessarily form a subsequence of T. While HMM assigns a joined probability $P(T, L) = \prod_i P(l_i|l_{i-1})P(T|l_i)$ to couples of sequences of observations $T = t_{1:n}$ and labels $L = l_{1:n}$, the CRF rather assigns a conditional probability $P(L|T) = \frac{1}{Z} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{F} \lambda_j f_j(l_{i-1}, l_i, t_{1:n}, i)\right)$ to them; where Z is the normalization factor and $f(\cdot)$ are the features functions that handle observation representations in CRF. However, after been trained on labeled text examples, both models are applied on an unlabeled text by using the Viterbi algorithm - this algorithm simply infers the label sequence having the highest probability. To help those models to get better results, some candidate descriptors of tokens have been designed. Our results related to the definition of some features, showing how they improve HMM and CRF performances are summarized in [3].

A lot of features have been defined; that are either based on the writing style of court decisions (e.g. line length, line words, neighbors words, word properties), or extended properties of tokens (e.g. part-of-speech tags, word topic). To optimize the feature sets, two feature selection algorithms have been tested for CRF: the bidirectional search (BDS) and Sequential Floating Forward Selection (SFFS).Those algorithms add or remove features over iterations as the results improve or not. Our actual experiments show that those algorithms are able to reduce significantly the quantity of features but only slightly improve the results. Their disadvantage is the duration of the selection phase - up to ten hours. We also studied four different segment representations (IO, BIO, IEO, and BIEO) that are different ways of tagging tokens [4]. BIO, IEO, and BIEO help to detect successive entities, as the beginning or the end of each entity is differently labeled. We also observed that those three representations actually improve results at the cost of a training time increase. Beside the segment representation and feature selections, labeling more data also obviously improve result within a limit. This suggests to correctly select the examples in order to cover the different variants of text structure.

3 A simple approach for claims extraction

During a case, each party expresses claims willing to be granted by the judges. Court decisions summarize, beside facts, claims and arguments of the parties, as

well as answers and reasons of judges. There are a lot of diverse styles to express claims and results: explicitly (in a dedicated paragraph precising the party to condemn, the corresponding fault, and eventually how much we claim), implicitly (need interpretation of texts), through reference to previous decisions (in appeal decisions, judges partially or totally confirm or reverse previous judgments). Hence the expression of claims is not standard and moreover, categories of claims are not all yet known. This makes claim extraction a difficult task. That is why our first approach is fully supervised and iterative, i.e. by trying to solve the problem considering the specificity of each category.

Firstly, a set of characteristic terms (n-grams) of a category is extracted by computing a score evaluating the likelihood for terms to occur in texts of that category. Several supervised global weighting methods have been tested: relative frequency [5], correlation coefficient [6], ... All these methods measure how much a term is relevant for a category by using labeled examples, i.e. *positive* texts (within the category) and *negative* ones (outside the category). Since, it may be hard to build a large set of labeled negative samples, we used a very large set of unlabeled data by assuming that the considered category only covers a small proportion of this large set. In practice, this postulate is true for the majority of the categories but fails for some categories covering a large subset of decisions (e.g. *article 700 du code de procédure civile*).

Secondly, to determine whether the decision refers to a category, we represent texts as vectors and then train a classifier on labeled examples. The dimensions of the vector are the terms learned from training sets that are the most relevant to the category; with the weight of a term $w*$ for a text t defined as:

$$weight(w*, t) = weight_{local}(w*, t) * weight_{global}(w*) * factor_{normalization}$$

The local weight is generally a term frequency TF-based method like term presence TP, or the logarithm of the term frequency. In our experiment, we have tried different configurations with different sizes of n-grams ($1 \leq n \leq 5$) fixed or not, global and local weighting methods, vector sizes, classifiers (SVM, Naive Bayesian, ...). Classifications on different categories have been tested (e.g. *dommages-intérêts pour procédure abusive, trouble de voisinage, article 700, prestation compensatoire*). Tests show that this classification task is actually straightforward since some configurations reach a 100 % accuracy even by using the simple Inverse Document Frequency (IDF) schema on all words. Indeed, very few terms can lead to very accurate classification since the vocabulary associated to the categories is very specific, short and common in justice.

However, our expectation was that the optimal subset of terms (for classification) would help locate the amount of money claimed and eventually granted - by evaluating distances between those terms and the mentioned quanta. Note however that the quantum can be mentioned before or after the terms - and sometimes the closest quantum to the terms is not the targeted one; in addition a quantum can also refer to a past decision or a claim. To tackle the problem of quantum/claim linking we assumed that it might help to zone the different contexts to further match claims and results using parties' names mentioned in

that zone. In the setting considered so far, the problem has been simplified by assuming that a decision contains at most one claim of a given category. By training a classifier on all the decisions, we expect to distinguish decisions with granted claims and decisions with rejected claims for a particular category of claim. Results on the datasets analyzed show the modest but interesting performance of the approach: maximal F1-measures at 75.52 for *article 700*, and at 86.66 for *trouble de voisinage*). To improve those results, instead of reducing features by selection, we are currently evaluating techniques of reduction by projection - in the style of the principal component analysis.

4 Conclusion

We have summarized our result on the study of sectioning and detection of legal named entities using probabilistic graphical models (HMM & CRF) - showing the interesting performance of those models in our application context. We are currently working on one of the more difficult tasks of the project: claims extraction. In particular, we are experimenting a simple method with which promising results have been obtained so far. As a schedule for the rest of the project, we also propose to study four additional tasks: (1) claims text segments detection using semantic and discourse analysis to deal with dependencies between statements; (2) comparing unsupervised and supervised claims categorization; (3) formalize as much as possible the information extracted so far in a knowledge base by applying disambiguation and resolution methods on extracted information; (4) analyze which factors and situations (contexts) are correlated with judges decision making - paving the way to predictive analysis.

References

1. Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
2. John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *International Conference on Machine Learning*, 2001.
3. Gildas Tagny Ngompé, Sébastien Harispe, Guillaume Zambrano, Jacky Montmain, and Stéphane Mussard. Reconnaissance de sections et d'entités dans les décisions de justice: application des modèles probabilistes HMM et CRF. In *In Extraction et Gestion des Connaissances - EGC 2017, Revue des Nouvelles Technologies de l'Information, Grenoble, France*, January 2017.
4. Michal Konkol and Miloslav Konopík. Segment representations in named entity recognition. In *International Conference on Text, Speech, and Dialogue*, pages 61–70. Springer, 2015.
5. Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):721–735, 2009.
6. Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *ACM SIGIR Forum*, volume 31, pages 67–73. ACM, 1997.