Predicting the Law Area and Decisions of French Supreme Court Cases

Octavia-Maria Şulea[1], Marcos Zampieri[2], Mihaela Vela[3], Josef van Genabith[3,4]
[1]University of Bucharest, Romania
[2]University of Wolverhampton, United Kingdom
[3]Saarland University, Germany
[4]German Research Center for Artificial Intelligence (DFKI), Germany
mary.octavia@gmail.com

Abstract

In this paper, we investigate the application of text classification methods to predict the law area and the decision of cases judged by the French Supreme Court. We also investigate the influence of the time period in which a ruling was made over the textual form of the case description and the extent to which it is necessary to mask the judge's motivation for a ruling to emulate a real-world test scenario. We report results of 96% f1 score in predicting a case ruling, 90% f1 in predicting the law area of a case, and 75.9% f1 score in estimating the time span when a ruling has been issued using a linear Support Vector Machine (SVM) classifier.

1 Introduction

Text classification methods have been used in a wide range of NLP tasks. This includes predicting information about authors of texts, such as age (Nguyen et al., 2013), gender (Rangel et al., 2013), personality (Sulea and Dichiu, 2015), and native language (Gebre et al., 2013), estimating the period in which a text was published (Niculae et al., 2014), the amount of subjectivity or sentiment expressed in texts (Balahur et al., 2014), and detecting pastiche (Dinu et al., 2012), plagiarism (Barrón-Cedeño et al., 2013), and influences from other authors (Ganascia et al., 2014). Classic machine learning algorithms such as Multinomial Naive Bayes and SVMs proved to be very reliable for these tasks, achieving high scores performance.

In this paper, we apply text classification methods to legal documents. We explore the use of bag of words (BOW) and linear SVM classifiers in predicting a case's ruling, law area, and the date in which a ruling was issued. We apply these methods to a large corpus of court rulings issued by the French Supreme Court with over 126,000 documents, spanning from the 1800s until the present day.

To the best of our knowledge, several NLP tasks have been carried out on legal texts, most notably text summarization (Farzindar and Lapalme, 2004; Galgani et al., 2012), however, as evidenced in Section 2, the use of text classification to predict court rulings is an under-explored area. The recent study by Aletras et al. (2016) on predicting decisions of the European Court of Human Rights (ECHR) is among the few examples of such attempts.

2 Related Work

In the legal domain, text classification has been more important to forensics (De Vel et al., 2001; Sumner et al., 2012; Pérez-Rosas and Mihalcea, 2015) than to predict information in legal texts such as case descriptions, rulings, and court decisions. General NLP methods, on the other hand, have played an important role in the intersection between artificial intelligence and law, a vibrant sub-area of research with international associations (e.g. IAAL[1]) and a number of specialized scientific conferences and workshops.

Palau and Moens (2009) investigate the extent to which one can automatically identify argumentative propositions in legal text, along with their argumentative function and structure. They use a corpus containing legal texts extracted from the European Court of Human Rights (ECHR) and classify argumentative vs. non-argumentative sentences with an accuracy of 80%.

Boella et al. (2011) present a classification approach to identify the relevant domain to which a

[1]http://www.iaail.org/

specific legal text belongs. Using TF-IDF weighting and Information Gain for feature selection and SVM for classification, reporting an f1-measure of 76% for the identification of the domains related to a legal text and 97.5% for the correct classification of a text into a specific domain.

The aforementioned studies by Farzindar and Lapalme (2004) and by Galgani et al. (2012) apply computational methods for the automatic summarization of legal texts. Such applications are developed to help law professionals in speeding up their work by providing shorter summaries of very long documents which are abundant in legal processes.

Studies applying text classification to legal documents include Hachey and Grover (2006), which proposed a system of classifying sentences for automatic court rulings summarization, and Gonçalves and Quaresma (2005), which used BOW, POS tags, and TF-IDF to classify legal text in 3,000 categories, based on a taxonomy of legal concepts, and reported 64% and 79% f1.

A few papers have been published on court ruling prediction. This includes the work by Katz et al. (2014), using extremely randomized trees, reporting 70% accuracy in predicting the US Supreme Court's behavior and, more recently, Aletras et al. (2016) proposed a computational method to predict decisions of the ECHR and reported their system's highest accuracy score as being 78%.

To the best of our knowledge, so far most work on predicting court rulings has been carried out on English data. No work has yet been carried out on French, such as the Supreme Court decisions we analyze in this paper. Moreover, to the best of our knowledge, previous work on court rule prediction did not take a temporal dimension into account and our work fills this gap.

3 Methods

3.1 Corpus

We use a diachronic collection of rulings from the French supreme court (*Court de Cassation*).[2] The complete collection[3] contains 131,830 documents, each consisting of a unique ruling and metadata formatted in XML. Common metadata available in most documents includes: law area, time stamp, case ruling (e.g. *cassation*, *rejet*, *non-lieu*, etc.), case description, and cited laws. In our supervised learning approach we use the metadata provided as 'natural' labels to be predicted by the machine learning system. In order to simulate realistic test scenarios, we remove all mentions from the training and test data that explicitly refer to our target prediction classes. In a pre-processing step we remove all surface forms of the words within the labels from the text data used to derive the predictive features.

All duplicate and incomplete entries in the dataset were excluded resulting in a corpus comprising 126,865 unique court rulings, each containing a case description and four different types of labels: a law area, the date of ruling, the case ruling itself, and a list of articles and laws cited within the description.

3.2 Tasks and Labels

In this section we present the process of defining labels in the dataset for the three tasks presented in this paper. The tasks and the respective section of the paper containing the results are summarized as follows:

1. Predicting the law area of a case (Section 4.1).
2. Predicting the court ruling based on the respective case description (Section 4.2).
3. Estimating when a case description and a ruling were issued (Section 4.3).

To reduce the feature and label space, we first removed accents and punctuation and lowercased all words in the description and ruling. Further preprocessing was needed to reduce the label space for each task. For task 1, we kept in the corpus all entries corresponding to the labels that had over 200 examples. This left us with 8 law area classes. Table 1 shows their distribution.

In establishing the ruling label set for predicting the case ruling (task 2), we were faced with a bigger challenge since, after the initial pre-processing, we were left with a list of 475 unique labels (from the initial 635). Looking at this list, we noticed that there were some entries which contained the same keyword repeated several times without having an overt interpretation for the repetition (e.g. *cassation partielle rejet rejet cassation* appeared 145 times in the dataset) as opposed to other multi-word labels which could

[2] https://www.courdecassation.fr/about_the_court_9256.html

[3] https://www.legifrance.gouv.fr

Law Area	# of cases
CHAMBRE_SOCIALE	33,139
CHAMBRE_CIVILE_1	20,838
CHAMBRE_CIVILE_2	19,772
CHAMBRE_CRIMINELLE	18,476
CHAMBRE_COMMERCIALE	18,339
CHAMBRE_CIVILE_3	15,095
ASSEMBLEE_PLENIERE	544
CHAMBRE_MIXTE	222

Table 1: Distribution of Law Area labels over the Case Descriptions

be easily interpreted (e.g *cassation partielle sans renvoi* which appeared 1,015 times).

An initial step, for better visualization of the ruling label space, was to do hierarchical clustering on the BOW occurrence vector representation for each label. We achieved this using Python's SciPy hierarchical functions with Ward distance (Figure 1). An immediate possibility of clustering the labels into 6-8 groups is apparent. We then investigate what might be the basis of this clustering and determined that keeping only the labels which had at least 200 examples was a good way to obtain this grouping.

On court ruling prediction, we carried out two sets of experiments. In the first one we considered only the first word within each label and only those labels which had over 200 entries in the corpus (first word setup). This lead to an initial set of 6 unique labels: *cassation*, *annulation*, *irrecevabilite*, *rejet*, *non-lieu*, and *qpc* (*question prioritaire de constitutionnalit*). The motivation behind using the first word, rather than using a more complex approach for the identification of the "correct" label, was based on the fact that in French the adjective follows the noun and that the labels consisted only of nouns, adjectives, and stop words.

In the second set of experiments, we considered all labels which had over 200 dataset entries and this time we did not reduce them to their first word. Table 2 shows the distribution of the ruling labels with over 200 examples each. Italics were used here to emphasize those labels which do not have an overt semantic interpretation. An important observation here is that, in the full, multi-word label extraction setup, *non-lieu* and *qpc*, which are known to be valid decisions of the French Supreme Court, are not selected as final labels, unlike in the first-word setup. This happens because they appear at the beginning of several rare labels (e.g. non-lieu a statuer, non-lieu a recevoir, qpc seule irrecevabilite, etc.). Therefore, there are not enough instances in the dataset with these labels for these labels to be selected. A similar phenomenon occurs with the rest of the labels when comparing the first-word to the multi-word setup.

First-word ruling	# of cases
rejet	68,516
cassation	53,813
irrecevabilite	2,737
qpc	409
annulation	377
non-lieu	246
Full ruling	**# of cases**
cassation	37,659
cassation sans renvoi	2,078
cassation partielle	9,543
cassation partielle sans renvoi	1,015
cassation partielle cassation	1,162
cassation partielle rejet cassation	906
rejet	67,981
irrecevabilite	2,376

Table 2: Distribution of Case Ruling labels over the Case Descriptions

Finally, for temporal text classification (task 3), we initially considered the decade of the ruling and the case description. The distribution is shown on Table 3, with the 1970 being the most prolific in cases.

Period	# of CR	Period	# of CR
1880s	1	1870s	8
1810s	2	1880s	10
1820s	2	1890s	8
1830s	1	1910s	2
1840s	4	1920s	17
1850s	9	1930s	29
1860s	9	1940s	15
1950s	84	1960s	4,797
1970s	23,964	1980s	18,233
1990s	16,693	2000s	12,577
2010s	4,541		

Table 3: Distribution of Ruling Date labels over the Case Descriptions

As discussed in Zampieri et al. (2016) the definition of time spans for supervised temporal text

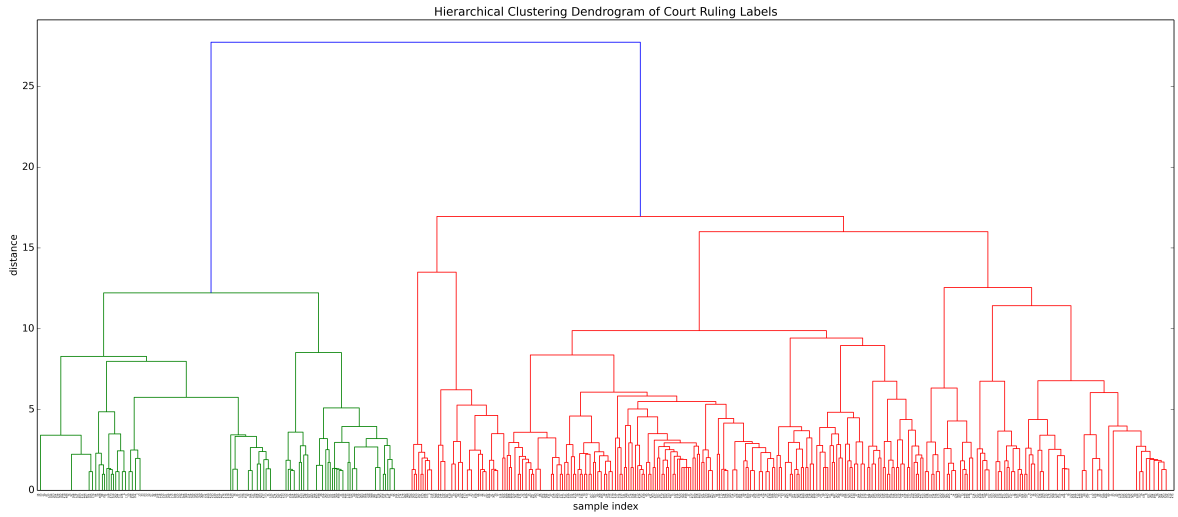

Figure 1: Dendrogram showing hierarchical clustering of ruling labels

classification is often arbitrary. Given that most cases were dated after 1960 and previous decades had only a few cases each, we divided the dataset into 7 classes by grouping all cases before 1960 under one label. Secondly, we considered fine-grained intervals by dividing the dataset into 14 classes merging classes before 1920 as follows: 1830-1840, 1850-1860, 1870-1880, 1890-1910.

3.3 Masking and Feature Selection

To make the three tasks more challenging and to emulate a real-world scenario, we had to eliminate the occurrence of each word of the label from the text of the corresponding case description.

For task 1, law area prediction, we eliminated all words contained in the respective label. For task 2, predicting the ruling, we initially eliminated from the case description all occurrences of the ruling word itself. We run ANOVA testing on the feature set (bag of words) and looked at the top 20 features to make sure that none of them could be construed as being directly linked to the label we were attempting to predict, so that a complete masking of the ruling was done within the case description text. In doing so, we realized the label was present both in its nominal form (e.g. *cassation, irrecevabilite*) and in its verbal forms (e.g. *casse, casser*). We eliminated these forms too. We finally investigated whether this technique of picking the top k classification features was good for identifying facts in the case description, aspects by which one would expect a lawyer to predict the judge's ruling. We did this by looking at the best 20 word bi-grams and tri-grams from the feature set. What we found instead were nouns with their articles (e.g. *la cause, le pourvoi*), prepositions with verbs and nouns (e.g. *pour etre, sur interpretation*), for bi-grams, and infinitival constructions (e.g. *et pour etre, occasion de faire*), for tri-grams.

Finally, for task 3, estimating the data of the case, we eliminate all digits from the case description. This has the disadvantage of removing digits that may refer to cited laws thus making the task even more challenging.

3.4 Computational Approach

We approach the tasks using a text classification system based on the scikit-learn implementation (Pedregosa et al., 2011) of the LIBLINEAR SVM classifier (Fan et al., 2008). As features, we investigate the capacity of word unigrams (bag of words) and bigrams (bag of bigrams) frequencies to capture the appropriate differences between case descriptions. We extract these features using scikit-learn's CountVectorizer.

Since these features rendered lower performance in temporal classification than in the first two, we also look at other features as proposed in Niculae et al. (2014) to improve the performance. Specifically, we couple BOW with the type-token ratio of each case description computed in the following way:

$$word\_type\_token = \frac{\#unique\_words}{\#total\_words}$$

As the dataset is imbalanced, we employ stratified 10-fold cross-validation for all experiments, since this validation method maintains the initial distribution over each fold. We compare our scores against a random baseline classifier implemented in scikit-learn as the DummyClassifier which takes into consideration the dataset's initial distribution. We report average precision, recall, and f1 scores over all labels. The C hyperparameter for the linear SVM was set to 0.1 in all experiments employing SVMs.

4 Results

In this section we report the results obtained for the three tasks: (1) predicting the law area of a case, (2) predicting the ruling of a case based on a case description, and (3) estimating the date of a case.

4.1 Law Area

In the first experiment, we apply the SVM classifier to predict the law area of a case. Table 4 shows the results of this classifier applied to 8 classes containing at least 200 instances each presented in Table 2.

Model	P	R	F1	Acc.
SVM	90.9%	90.2%	90.3%	90.2%
baseline	17.7%	17.7%	17.7%	17.%

Table 4: Classification results for the law area prediction task using Linear SVM on 8 classes

The results show that on average our system is able to predict the law area of a case and court ruling with high precision, recall, and f1 score, well above those of the random baseline.

4.2 Court Ruling

In this section we present the results obtained in the second task, ruling prediction based on a case description. The results are presented in Table 5. We report the scores of the experiments when run on the first-word (6 classes) as well as multi-word setups (8 classes) for label extraction discussed in Section 3.2.

We observe an apparent 6 percentage points decrease in average scores when the classifier is trained on the dataset with more classes. This is in tune with the characteristic of classifiers such as SVM which suffer from imbalanced data and is to a certain extent expected since the class imbalance is significant. However, it is important to note that the drop is only apparent, since the increase in number of classes leads to a decrease in the random baseline performance and thus the difference between the baseline scores and our method actually grows by 4 percentage points from the first-word setup.

Model	P	R	F1	Acc.
6 cls SVM	97.1%	96.9%	97.0%	96.9%
6 cls baseline	47.7%	47.7%	47.7%	47.7%
8 cls SVM	93.2%	92.8%	92.7%	92.8%
8 cls baseline	40.6%	40.6%	40.6%	40.6%

Table 5: Classification results for the ruling prediction task using Linear SVM

In terms of previous work, unfortunately a systematic and thorough comparison with Katz et al. (2014) and Wongchaisuwat et al. (2016) is not possible since we are not using the same corpus nor working on the same language as these two papers. Even so, our method appears to surpass both, in terms of f1 score, in predicting the ruling of a court, based on previous examples. One main difference might be the judicial system which is known to be more predictable (offering the judges less interpretation freedom) in the case of the French Supreme Court.

4.3 Temporal Classification

For the third task, estimating the date of case and ruling, we use the same approach as previous experiments, a linear SVM classifier trained on bag of unigrams and bag of bigrams as features. Results in two settings, one containing 7 classes and the other containing 14 classes, are reported in Table 6

The general tendency of traditional supervised classification algorithms is to increase their performance as the number of classes or imbalance between classes decreases. Our experiments show that we manage to preserve the difference between the baseline performance and that of our system on different tasks (ruling prediction and temporal classification), with varying number of classes and initial distributions, which suggests that these techniques are robust for our purpose. However, from a user perspective, where error rate needs to be low, we expect this observation to not be useful and we therefore also run the SVM experiments with type-token ratios as features. On their own,

Subtask	Model	Precision	Recall	F1	Accuracy
7-class	SVM 1-gram	69.9%	68.3%	68.2%	68.3%
7-class	SVM 2-gram	**75.9%**	**74.3%**	**73.2%**	**74.3%**
7-class	baseline	19.2%	19.2%	19.2%	19.2%
14-class	SVM 1-gram	69.1%	68.6%	68.5%	68.6%
14-class	SVM 2-gram	**75.6%**	**74.2%**	**73.9%**	**74.2%**
14-class	baseline	19.1%	19.1%	19.1%	19.1%

Table 6: Classification results for temporal prediction using Linear SVM

they were able to reach a little above the random baseline (43% f1 vs. 19% for the random). Interestingly, type-token ratio did not increase the performance of the classifier when combined with BOW.

5 Conclusions and Future Work

In this paper we investigated the application of text classification methods to legal texts from the French Supreme Court. To the best of our knowledge, this is the first work to: (1) apply text classification to predict the rulings on a French dataset, (2) carry out temporal text classification experiments on legal texts. The paper also reports high performance in the task of predicting court rulings.

We showed that a linear SVM classifier trained on BOW can obtain high f1 scores in predicting the law area and the ruling of a case, given the case description. Estimating the date of cases turned out to be more difficult to learn using bag of words and lexical richness features (type-token ratio), but this may be due to the highly imbalanced dataset (i.e. too few examples from the minority classes) or to the possible fact that the language used by judges of the French Supreme Court over the years has not changed much. This final observation is worth further investigation.

We also looked at ways of masking the case description to convey as little information as possible regarding the ruling itself making the task more challenging. This method showed that the word bigrams and trigrams deemed to be the most salient in predicting the ruling are not actually tied to any factual information particular to one case, but more related to formulaic expressions typical for a particular ruling. In future work, we would like to extend this investigation to the sentence level and see if the sentences that are considered most effective in predicting the ruling are of factual nature.

Our work is proof of concept that text classification techniques can indeed be used to provide valuable assistive technology base as support for law professionals in obtaining guidance and orientation from large corpora of previous court rulings. In the future, we would like to investigate the extent to which a more accurate draft form can be induced from the court's case description.

Acknowledgements

This work was carried out while the first and the second author, Octavia-Maria Şulea and Marcos Zampieri, were at the German Research Center for Artificial Intelligence (DFKI).

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Computer Science* 2:e93.

Alexandra Balahur, Rada Mihalcea, and Andrés Montoyo. 2014. Computational Approaches to Subjectivity and Sentiment Analysis: Present and Envisaged Methods and Applications. *Computer Speech & Language* 28(1):1–6.

Alberto Barrón-Cedeño, Marta Vila, M Antònia Martí, and Paolo Rosso. 2013. Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics* 39(4):917–947.

Guido Boella, Luigi Di Caro, , and Llio Humphreys. 2011. Using Classification to Support Legal Knowledge Engineers in the Eunomos Legal Document Management System. In *Proceedings of JURISIN*.

Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining E-mail Content for Author Identification Forensics. *ACM Sigmod Record* 30(4):55–64.

Liviu P Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2012. Pastiche Detection Based on Stopword Rankings: Exposing Impersonators of a Romanian Writer. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9:1871–1874.

Atefeh Farzindar and Guy Lapalme. 2004. Legal Text Summarization by Exploration of the Thematic Structures and Argumentative Roles. *Proceedings of the Text Summarization Branches Out Workshop* .

Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining Different Summarization Techniques for Legal Text. In *Proceedings of the Hybrid Workshop*.

Jean-Gabriel Ganascia, Pierre Glaudes, and Andrea Del Lungo. 2014. Automatic detection of reuses and citations in literary texts. *Digital Scholarship in the Humanities* 29(3).

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving Native Language Identification with TF-IDF Weighting. In *Proceedings of the BEA Workshop*.

Teresa Gonçalves and Paulo Quaresma. 2005. Evaluating Preprocessing Techniques in a Text Classification Problem. In *Proceedings of the Conference of the Brazilian Computer Society*.

Ben Hachey and Claire Grover. 2006. Extractive Summarisation of Legal Texts. *Artificial Intelligence and Law* 14(4):305–345.

Daniel Martin Katz, Michael J. Bommarito II, and Josh Blackman. 2014. Predicting the Behavior of the Supreme Court of the United States: A General Approach. *CoRR* abs/1407.6333.

Dong-Phuong Nguyen, Rilana Gravel, RB Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A Study of Language and Age in Twitter. In *Proceedings of ICWSM*.

Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. 2014. Temporal Text Ranking and Automatic Dating of Texts. *Proceedings of EACL* .

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of the ICAIL*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in Open Domain Deception Detection. In *Proceedings of EMNLP*.

Francisco Rangel, Efstathios Stamatatos, Moshe Moshe Koppel, Giacomo Inches, and Paolo Rosso. 2013. Overview of the Author Profiling task at PAN 2013. In *Proceedings of CLEF*.

Octavia-Maria Sulea and Daniel Dichiu. 2015. Automatic Profiling of Twitter Users Based on Their Tweets: Notebook for PAN at CLEF 2015. In *Proceedings of CLEF*.

Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J. Park. 2012. Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In *Proceedings of ICMLA*.

Papis Wongchaisuwat, Diego Klabjan, and John O McGinnis. 2016. Predicting Litigation Likelihood and Time to Litigation for Patents. *arXiv preprint arXiv:1603.07394* .

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016. Modeling Language Change in Historical Corpora: The Case of Portuguese. In *Proceedings of LREC*.