LDSI - Literature survey

Name: Mahaut Gérard
Gloss ID: <tum_ldsi_21>

In this literature survey, based on six papers carefully selected [1, 2, 3, 4, 5, 6], we go through the four basic steps of NLP pipelines. We look at the datasets, the word embeddings techniques, the classification models and the evaluation methods used. We finish the survey by a discussion focusing on correlation between type of sentences and classification results. We focus on legal texts' sentence classification. In the literature, we note that legal sentence's classes are mainly referred to as rhetorical roles, we also use this denomination in our survey. The purposes of legal texts' sentences classification are automatization of legal texts annotations, needed for supervised learning models training, and more generally support the work of law experts.

Data

Dataset [1, 3] consider Veteran Claims, 6153 sentences, issued by the U.S. Board of Veterans' Appeal (BVA). The sentence's types of [1, 3] are Finding, Reasoning, Evidence, Legal rule, Citation. This is especially interesting as we also use decisions from the BVA with similar sentence's types in our experiments. [2] uses German legal texts from German civil code (601 sentences) and from rental agreements (481 sentences). We note that the two types of documents used, yet legal texts, are of different types. [4] works with 592 sentences and 62 lists from Dutch laws. [5] considers judgements from the House of Lords (HOLJ corpus) and gathers 98645 sentences. This last paper has significantly more sentences than the other we surveyed. Finally, [6] uses legal texts from the Supreme Court of India (9308 sentences) and from the Supreme Court of the UK (18155 sentences). The two corpora are used to train independent models and then compared to study portability of models inside the legal domain. Papers either use existing legal corpora or create their own corpus by annotating legal documents [2, 6]. Manual annotations are done by legal domain experts and [2] uses the Gloss tool.

Pre-processing The pre-processing steps include different normalization steps: headings sentence removal [3], numeric strings handling (removal [3], grouping [1, 4], spelling [2]), non-alphanumeric characters (such as punctuation) removal [2, 3], stemmer (NLTK Snowball) [3, 4], lemmatization [2], stop words removal [2, 4], lower case [1, 4], removal of words below the minimal term frequency [4], removal of lines break [2], removal of duplicated whitespaces [2]. It's interesting to note that [2, 4] analyze the efficiency of pre-processing methods by not using them all at the same time and comparing the performance for different sets of pre-processing steps. [4] found that the better set is using binary weight, stop list words and minimum term frequency of 2. [2] presents lemmatization as the central pre-processing step to integrate. Finally, sentence segmentation is also used as a pre-processing step. The spaCy2 package is found useful by [2, 6].

Word embeddings

There are two kinds of word embedding methods used in the surveyed papers. The first type is pre-trained models for word embeddings such as GloVe [1], Fasttext [1], Law2Vec [1, 6], Word2Vec [2], and Google News [6]. Some of the pre-trained models are law specific (Law2Vec), they tend to perform better than models pre-trained on general corpora (Google

News) [6]. The second type is the bag-of-word approach [2, 4] such as TFIDF and binary featurization, and the vectorizer approach [3] considering individual tokens, bigrams and trigrams. It is interesting to note that the paper that compares both a pre-trained model and a bag-of-words approach [2] concludes that the bag-of-word approach outperforms the pre-trained model featurization. Best embeddings vary from one paper to the other and the results from [6] highlight that it depends on the type of legal text used.

Models

Types of Models Lots of different classification models are considered, there are all supervised approaches: sentence annotation is needed in the first place. First, there are the linear models: Naïve Bayes [3, 5], Multinomial Naïve Bayes [2], Logistic Regression [2, 3], Decision Tree [2, 5], and Support Vector Machines (SVM) [2, 3, 4]. Second, there are the nonlinear models: Random Forest [2], Multilayer perceptron [2], polynomial SVM [5]. Finally, there are the deep learning models based on LSTM (bidirectional LSTM, CNN-LSTM, Hier-BiLSTM), Transformers (Tf-BiLSTM), Statistical models (BiGRU-CRF), and Attention based networks and combination of them are tested by [1, 6]. It is also interesting to note that [4] tests a knowledge engineering (KE) method for classification.

k-folds cross validation is used by half of the papers 10 folds for [2], 5 folds for [6] and Leave-One-Out for [4]. This approach aims at handling overfitting caused by small datasets.

Best Models The best performing models differ from one paper to the other. [1] concludes that Bi-LSTM is the best performing model with 91% of accuracy and 87% of f1-score. [2] presents that the best configuration uses a bag-of-word approach and a SVM classifier or Decision Tree classifier, therefore linear models. They have a f1-score of 83%. [3] finds that SVM and Logistic Regression are the best performing models with f1-scores around 79%. [4] argues that the KE method performs similarly (~ 90% accuracy) than SVM, and therefore the KE approach is to be preferred. [5] concludes that the Decision Tree classifier is the best performing model with around 60% f1-score. Finally, [6] proposes BiLSTM-CRF with a f1-score of 82% for the Indian Court texts and 60% for the UK Court texts.

In short, from the papers surveyed outside of deep learning methods, complex (non) linear models are not a solution towards achieving better legal sentence classification as simple models often perform better. However, deep learning models are presented as very efficient. As legal texts are from various types and jurisdictions, and as the classes differ from one paper to the other, we should also keep in mind that comparing scores from one paper to the other is often not relevant. It highlights the importance of baseline comparison inside each paper and of consideration of several evaluation scores. In addition, [6] notes that there is often a trade-off between performance (with Deep learning) and explainability (simple machine learning). The compromise to adopt is then case specific. Finally, we want to note that the two papers focusing on deep learning methods [1, 6] are the most recent ones (2020 and 2021). Deep learning methods for legal sentence classification seem to be the future towards better classification performance.

Evaluation

The evaluation of classification models is an important part when building a solution for sentence classification. Different kinds of scores are reported in all the papers: accuracy, recall, precision, f1-score, micro-averaged and macro-averaged, and confusion matrix. The scores are computed between the gold standard (human annotation) and the predicted

labels (by machine learning models). Most of the papers stress that the human annotations are never perfect, and that especially for overlapping or subjective classes the human annotations fail to produce ground truth classification [6].

Further discussion - What are the rhetorical types that are the most difficult/easy to classify?

Overall, types such as Citation / Reference / Quotations [1, 2, 3, 5], Evidence [1, 3], and Legal Rule [3] have high scores. These are types that are easy to spot both for human and machine learning models. In addition, a lot of sentences have these types in legal texts.

On the other hand, Reasoning [1, 3], Argument [6], Cue Phrases [5] sentences perform badly. As we see in the confusion matrix of [4], Permission, Obligation, and Delegation types are mixed, this leads to relatively bad results. [6] summarizes two reasons for poor classification results of some classes. First, some classes appear to be subjective and human domain expert annotators also struggle to classify them or to agree on an annotation. Second, the lack of data representing certain classes. Under-represented classes tend to have poor classification results because machine learning models fail to capture the underlying representation of them. This is in fact a common problem in the AI field, referred to as unbalanced datasets, also highlighted by [2].

References

[1] S. R. Ahmad, D. Harris and I. Sahibzada, "Understanding Legal Documents: Classification of Rhetorical Role of Sentences Using Deep Learning and Natural Language Processing," *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, (2020).

[2] Glaser, Ingo et al. “Classifying Semantic Types of Legal Sentences: Portability of Machine Learning Models.” *JURIX* (2018).

[3] Walker, Vern R. et al. “Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning.” *ASAIL@ICAIL* (2019).

[4] E. de Maat, K. Krabben, and R. Winkels. “Machine Learning versus Knowledge Based Classification of Legal Texts.” *JURIX (2010).*

[5] B. Hachey , and C. Grover “Sentence Classification Experiments for Legal Text Summarisation” *JURIX (*2004)

[6] Bhattacharya, P., Paul, S., Ghosh, K. *et al.* DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artif Intell Law* (2021)