Categorizing Roles of Legal Texts via Sequence Tagging on Domain-Specific Language Models

Sourav Dutta

Huawei Ireland Research Centre, Dublin, Ireland

Abstract

Automatically understanding *rhetorical roles* of text snippets within a legal document provides an interesting problem, enabling several downstream tasks like summarization of legal judgments, similar legal text search, and case analysis. The task is challenging as legal case documents are domain-specific, usually not well-structured, and rhetorical roles may be subjective.

To this end, we present how sentence embeddings from *domain-specific pre-trained language model* can be combined with a *sequence tagging classifier*, to better understand the implicit sections within legal documents via long-term relationships, for sentence classification. Our proposed methodology secured the **1st** rank, with an F1 score of 0.557 on the shared task 1 in the "Artificial Intelligence for Legal Assistance" (AILA) track of the *Forum of Information Retrieval Evaluation* (FIRE), 2021.

Keywords

Legal Data Analytics, Rhetorical Role Labelling, Sentence Classification, Language Model

1. Introduction

The legal framework in most countries rely on two primary sources – *Statues* and *Precedents*. *Statutes* are bodies of written law, such as the Constitution of a country, while *Precedents* denote the prior cases as decided in other Courts of law. A legal representative, when presenting a case in a court of law, must adhere to facts, relevant precedents and statues. Legal documents tend to large and majorly unstructured [1, 2], necessitating novel systems for automatically understand and segment such documents into coherent meaningful parts. Such frameworks would improve readability and assist legal representatives, but also enable diverse downstream tasks such as semantic case search [3], legal document summarization [4, 5], and legal analysis [6].

Rhetorical role labelling of sentences in a legal document refers to understanding the semantic function a sentence in the document [7]. Although, legal case documents, in general, follow a common thematic structure with implicit sections like "Facts", "Issues" and "Arguments given by parties", this information is generally not specified explicitly in free-flowing case documents and various themes often interleave with each other. To alleviate the above challenges, current research in the domain of legal informatics involves the use of machine learning approaches for supervised classification of legal texts. However, the presence of limited training data and domain-specificity provides a challenge for automating the identification of rhetorical roles.

Forum for Information Retrieval Evaluation, December 13–17, 2021, India
sourav.dutta2@huawei.com (S. Dutta)
0000-0002-8934-9166 (S. Dutta)

To this end, we propose a framework for rhetorical role labelling of legal sentences. To provide domain information, we rely on sentence embeddings from a fine-tuned domain-specific transformer based language model. This enables the contextual understanding of the domain-specific vocabulary and their relationship with the rhetoric labels. Further, as documents implicitly have an underlying thematic structure, we adapt a Gated Recurrent Unit-Conditional Random Field (GRU-CRF) based sequence tagging classifier to categorize the sentences into the rhetoric labels. Our proposed methodology secured the **1st** rank, with an F1 score of 0.557 on the shared task 1 in the "Artificial Intelligence for Legal Assistance" (AILA) track of the *Forum of Information Retrieval Evaluation* (FIRE), 2021[8, 9]. Variants of the classification framework were also ranked at 2nd and 3rd positions – depicting the effectiveness of our framework.

1.1. Related Work

Legal document analysis and role labelling have traditionally been an expensive manual process by domain experts. With the advent of legal analytics and machine learning techniques, automatic labelling of the rhetorical role of legal sentences was studied. However, such approaches rely on the availability of manual annotated datasets based on a set of rules crafted from domain knowledge. An in-depth annotation study and curation of a gold standard corpus for the task of sentence labelling can be found in [10].

There have been several prior attempts towards automatically identifying rhetorical roles of sentences in legal documents. A method for identification of factual and non-factual sentences was developed in [3] using fastText classifier, while Conditional Random Fields (CRF) were used for rhetorical role labelling in [4]. The use of rule-based scripts, requiring lesser training data, with machine learning approaches for rhetorical role identification was studied in [11].

Deep learning model like hierarchical BiLSTM-CRF based classifier was recently shown to perform better at than using only handcrafted linguistic features [7]. Further, use of pre-trained language models like RoBERTa [12], along with TF-IDF based semantic features [13] were shown to perform well on such tasks in AILA-2020.

2. Task Description

The task of rhetorical role labelling of sentences in legal case judgements was first introduced in AILA-2020 [14], following which a similar shared task was set up for AILA-2021 [8, 9]. The dataset was created based on legal judgments from the Supreme Court of India, and were subsequently manually labelled by legal experts [7].

The sentences of the documents are to be classified into one the following *seven* labels:

- **Facts**: sentences that denote the chronology of events that led to filing the case;
- **Argument**: sentences that denote the arguments of the contending parties;
- **Statute**: relevant statute cited;
- **Precedent**: relevant precedent cited;
- **Ruling by Lower Court**: sentences correspond to the ruling/decision given by the lower courts (e.g., Tribunal, High Court, etc.), as the dataset contains cases presented to the Indian Supreme Court with a preliminary ruling from the lower courts;

- **Ratio of the Decision**: sentences that denote the rationale/reasoning given by the Supreme Court for the final judgement; and
- **Ruling by Present Court**: sentences that denote the final decision given by the Supreme Court for that case document.

Overall, there were 70 annotated legal documents for training, with documents of differing lengths and the annotated classes were unbalanced. In total there were around 11K training sentences with *Ratio of the Decision* and *Facts* constituting 50% of the labels. The test suite contained 10 documents for submission, with around 850 sentences for classification.

3. Proposed Framework

In this section, we introduce the different modules of our proposed framework for the above task. In summary, we adopted 4 main strategies:

1. **Sentence Embedding** – Each sentence is encoded into a dense vector representation using sentence embedding obtained from a fine-tuned domain-specific language model. This provides semantic and contextual information of each sentence for classification tasks. We also explore the use of both generic as well as domain-specific sentence embedding techniques as discussed later;
2. **Structural Encoding** – Sentences belonging to a certain class might have structural properties that are similar to each other (while being different from the sentences of another class). For example, arguments might have different linguistic structure as compared to court rulings, in terms of word ordering and their dependencies. Hence, in a variant we employ sentence embeddings obtained from their dependency parse trees;
3. **Meta-Embedding** – Diverse embedding obtained from different encoding architectures when concatenated have been shown in the literature to increase the overall classification performance, as compared to the accuracy of individual embeddings. In our framework, we concatenate sentence embeddings obtained from different language models to form the final sentence representations provided to the classification layer; and
4. **Document-Sentence Sequence Classification** – As opposed to classifying each sentence into the provided 7 labels, to incorporate a global view on the classification task, we adapt sequence tagging [15] for sentence classification by considering long-term label dependency between the sentences within a document. The intuition is that, documents inherently might have a logical structure, which if considered, should improve the overall accuracy of classification. For example, a document might first state the facts and arguments, followed by ruling of the lower court, and finally the current court ruling.

We next discuss the different strategies in more details, along with the various models developed for the *rhetoric role labelling* task.

3.1. Sentence Embeddings

We compute high-dimension embeddings of sentences of the legal documents by using the following two language models:

- *Domain-Specific Model*: To obtain domain based embedding of sentences, we use the pre-trained *"legal-bert-base-uncased"* language model [1], which is pre-trained on different legal documents. This enables our classification model to understand domain-based terminology and semantic contextual information related to the domain of the task. We further fine-tune the above model using the training data (of the task) with a batch size of 64 and learning rate set to $2e-5$.
- *Generic Model*: For inducing a generic semantic understanding of natural language text (for understanding general meaning of the sentences), we use the pre-trained sentence transformer model *"all-mpnet-base-v1"* [2]. This model is also fine-tuned on the training dataset provided, with the same parameters as mentioned above (for the domain-specific model).

For both the above models, we use mean-pooling strategy to obtain the sentence embeddings.

3.2. Structural Encoding

This module helps our framework to learn encodings of sentence structures, as additional information cues. As mentioned previously, certain sentences (like arguments) might have specific linguistic structures that might help in their classification. To this end, we use an unsupervised approach based on *Graph2Vec* [3]. The dependency parse tree (obtained using SpaCy software) of each sentence is converted to an undirected graph and fed to Graph2Vec to get the embedding of the sentence structures. Specifically we use the following parameters for Graph2Vec: wl-iterations=1, dimensions=512, epochs=20, and min-count=1.

3.3. Meta-Embedding

We combine the various sentence embeddings obtained from the above modules – *domain-specific, generic, and structural representations* – in different combinations for the variants of our framework (specified later). Meta-embedding is constructed by simply *concatenating* the different sentence embeddings (the order of concatenation is irrelevant).

3.4. Document-Sentence Sequence Classification

The final classification module generates the output classes for each of the input sentences. As discussed above, we deviate from the approach of individual sentence classification, but look at the sentences in unison within a document – as this might enable the learning of implicit structure and sentence class dependency within a document (e.g., a fact might be followed by argument and then court ruling).

Sequence tagging classifiers have been used in the literature for Named-Entity Recognition (NER), Part-of-Speech (POS) tagging, and chunking [15]. We adopt a similar approach but for document-sentence level classification. Specifically, in NER task, sentences are composed of words, and each word in the word sequence (i.e., the sentences) is classified into entity types.

[1] available at huggingface.co/nlpaueb/legal-bert-base-uncased
[2] available at huggingface.co/sentence-transformers/all-mpnet-base-v1
[3] karateclub.readthedocs.io/en/latest/modules/root.html#karateclub.graph_embedding.graph2vec.Graph2Vec

Similarly, we consider a document to consist of sentences, and each sentence in the sentence sequence (i.e., the document) is categorized by our classification layer. this enables information flow across the sentences in the document for identifying the sentence classes, improving the overall classification accuracy of the proposed framework.

Specifically, we use a bi-directional Gated Recurrent Unit (GRU) [16] along with a Conditional Random Field (CRF) [17] layer on top as the classification module. A document is represented as a sequence of sentences, and the input to the GRU units are sentence embedding obtained from the different models as discussed above. We also use a fully connected and a dropout layer for classification with Adam optimizer. We also use class weights during training the classifier layer to account for training class data imbalance. The implementation uses tensorflow-keras and tf2crf libraries, for classifying the sentences into the 7 categories.

4. AILA-2021 Task 1 Results

In this section, we report the performance of our framework on the rhetoric role labelling task of AILA-2021. We consider the following *three* variants of our framework – where the input sentence embeddings are different and depends on the embedding techniques used to create the sentence meta-embedding. We use the following setup for the variants, as:

- **Variant 1** — Only the fine-tuned domain-specific embedding (from legal-bert-base-uncased language model) for each of the sentences is provided as input to the classification layer. The total sentence embedding dimension is 768.
- **Variant 2** – Here the fine-tuned domain-specific embedding along with unsupervised structural embedding (from legal-bert-base-uncased and Graph2Vec) is concatenated for each of the sentences is fed as input to the classification layer. That is, sentence meta-embedding (fine-tuned legal-bert-base-uncased + unsupervised graph2vec embedding) is obtained from the 2 methods. The total sentence embedding dimension is 1280.
- **Variant 3** – In the final run, we concatenate embeddings from all the three models to obtain overall sentence meta-embedding, which is passed as input to the classification layer. That is, we use domain-specific, generic and structural embedding to create the sentence embedding (i.e, fine-tuned legal-bert-base-uncased + unsupervised graph2vec + fine-tuned all-mpnet-base-v1). The total sentence embedding dimension here is 2048.

The classification layer and other parameters are kept the same across all the different variants.

Empirical Scores. The classification performance is computed in terms of *Precision*, *Recall* and *F1-Score* within each classification group, and is averaged to obtain the *macro-F1 score* for the entire task. The results obtained by the proposed framework in the different category classes are tabulated in Table 1.

We obtained the **1st**, **2nd** and **3rd leaderboard positions** for the task on the overall performance of our framework, with a best F1 score of 0.557 from variant 1. It can be observed that using only the fine-tuned language model for obtaining sentence embeddings achieved the best overall results on this task dataset. The addition of structural embedding information also produced comparable results (with a slight decrease in precision, but increase in the recall

Table 1
AILA-2021 Task 1: Sentence Rhetorical Role Labelling Performance Scores.

Framework / Category	Variant 1			Variant 2			Variant 3		
	Precision	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*	*Precision*	*Recall*	*F1 Score*
Overall	**0.548**	0.616	**0.557**	0.528	**0.619**	0.551	0.511	0.627	0.549
Facts	0.765	0.695	0.728	0.749	0.749	**0.749**	0.676	0.724	0.699
Argument	0.808	0.539	0.646	0.767	0.590	0.667	0.735	0.641	**0.685**
Statute	0.619	0.867	**0.722**	0.571	0.800	0.667	0.483	0.933	0.636
Precedent	0.296	0.746	**0.424**	0.270	0.612	0.374	0.316	0.448	0.370
Ruling by Lower Court	0.000	0.000	0.000	0.000	0.000	0.000	0.111	0.133	**0.121**
Ratio of the Decision	0.751	0.620	0.679	0.768	0.620	**0.686**	0.737	0.627	0.677
Ruling by Present Court	0.595	0.846	0.698	0.568	0.962	**0.714**	0.523	0.885	0.657

scores). It is interesting to observe that the inclusion of generic sentence embeddings degraded the classification performance of our model – as such embeddings fail to generalize in the presence of domain-specific vocabulary.

An important observation is that the *Ruling by Lower Court* is the hardest class for our framework. This can be attributed to the presence of the closely related *Ruling by Present Court* class, which might be confusing the classification module. This can be validated by the high recall score but low precision score of the *Ruling by Present Court* category (see last row of Table 1) – which indicates the model probably classifies both types into this single category.

5. Conclusion

In the work we present a framework for automatically assigning rhetorical roles to sentences in legal documents. Our approach relies on embeddings from fine-tuned domain-specific transformer based language model and sentence structure. The use of sequence tagging based GRU-CRF classification layer enabled us to model long-distance sentence relationships within documents for better classification performance. We secured ranks 1, 2 and 3 results on the AILA-2021 shared task 1 with a best overall F1-score of 0.557.

References

[1] O. Shulayeva, A. Siddharthan, A. Z. Wyner, Recognizing Cited Facts and Principles in Legal Judgements, Artificial Intelligence and Law 25 (2017) 107–126.

[2] P. Bhattacharya, K. Hiware, S. Rajgaria, N. Pochhi, K. Ghosh, S. Ghosh, A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments, in: ECIR, 2019.

[3] I. Nejadghoii, R. Bougueng, S. Witherspoon, A Semi-supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases, in: JURIX, 2017.

[4] M. Saravanan, B. Ravindran, S. Raman, Automatic Identification of Rhetorical Roles using Conditional Random Fields for Legal Document Summarization, in: International Joint Conference on Natural Language Processing, 2008.

[5] A. Farzindar, G. Lapalme, LeTSum, An Automatic Legal Text Summarizing System, 2004.

[6] J. Savelka, K. D. Ashley, Segmenting U.S. Court Decisions into Functional and Issue Specific Parts, in: JURIX, 2018.
[7] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, A. Wyner, Identification of Rhetorical Roles of Sentences in Indian Legal Judgments, in: Proc. International Conference on Legal Knowledge and Information Systems (JURIX), 2019.
[8] V. Parikh, U. Bhattacharya, P. Mehta, B. A., P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, Overview of the Third Shared Task on Artificial Intelligence for Legal Assistance at Fire 2021, in: FIRE (Working Notes), 2021.
[9] V. Parikh, U. Bhattacharya, P. Mehta, B. A., P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, FIRE 2021 AILA track: Artificial Intelligence for Legal Assistance, in: Proceedings of the 13th Forum for Information Retrieval Evaluation, 2021.
[10] A. Z. Wyner, W. Peters, D. Katz, A Case Study on Legal Case Annotation, in: JURIX, 2013.
[11] V. R. Walker, K. Pillaipakkamnatt, A. M. Davidson, M. Linares, D. J. Pesce, Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning, in: Workshop on Automated Semantic Analysis of Information in Legal Texts (with ICAIL), 2019.
[12] S. B. Majumder, D. Das, Rhetorical Role Labelling for Legal Judgements Using ROBERTA, in: Forum for Information Retrieval Evaluation-AILA, 2020.
[13] J. Gao, H. Ning, Z. Han, L. Kong, H. Qi, Legal Text Classification Model based on Text Statistical Features and Deep Semantic Features, in: Forum for Information Retrieval Evaluation-AILA, 2020.
[14] P. Bhattacharya, P. Mehta, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, Overview of the FIRE 2020 AILA Track: Artificial Intelligence for Legal Assistance, in: Forum for Information Retrieval Evaluation, 2020.
[15] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, arXiv:1508.01991, 2015.
[16] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.
[17] J. Lafferty, A. McCallum, F. C. N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: International Conference on Machine Learning (ICML), 2001, pp. 282–289.