Spectre@AILA-FIRE2020: Supervised Rhetorical Role Labeling for Legal Judgments using Transformers

Racchit Jain, Abhishek Agarwal and Yashvardhan Sharma

Department of Computer Science and Information Systems, Birla Institute of Technology and Science Pilani, Pilani Campus

Abstract

This paper presents the methodologies implemented while performing multi-class text classification for rhetorical role labeling of legal judgments for task 2 of the track 'Artificial Intelligence for Legal Assistance' proposed by the Forum of Information Retrieval Evaluation in 2020. A transformer-based language model (RoBERTa) pretrained over an extensive English language corpus and fine-tuned on legal judgments from the Supreme Court is presented along with its evaluation on standard classification metrics, precision, recall, and F-score.

Keywords

text classification, transformers, RoBERTa, rhetorical role labeling, AILA,

1. Introduction

The meaning of the phrase "Rhetorical Role labeling" in the legal context corresponds to identifying the semantic function a sentence serves in the legal document. For example, in legal case documents, there are facts, precedents, rulings by the court, etc. that form the document's semantic structure. With the rapid increase in the digitization of legal documents, automating the detection of rhetorical roles of sentences in a legal case document can allow easy summarizing, organization of case statements, case analysis, etc. Due to the high specificity and technical jargon in legal documents, rhetorical role labeling is an extremely challenging NLP task. Previous work has focused on using handcrafted features and linguistic approaches towards this classification problem. Bhattacharya et al.[1] introduced a neural model (Hierarchical BiLSTM CRF Classifier) which outperformed the previous approaches by a large margin by making use of pretrained legal embeddings.

The 'Artificial Intelligence for Legal Assistance' track proposed by FIRE 2020[2], comprised of two tasks. This paper will discuss task-2 of this track, 'Rhetorical Role Labeling for Legal Judgments'. Each team was provided with an annotated dataset of 50 Supreme Court legal documents. Every sentence in the document was given one out of the seven labels: {Facts, Ruling by Lower Court, Argument, Statute, Precedent, Ratio of the decision, Ruling by Present Court}. The presented approach achieved an overall 9th rank among all the runs according to

Forum for Information Retrieval Evaluation 2020, December 16–20, 2020, Hyderabad, India

f20190145@pilani.bits-pilani.ac.in (R. Jain); f20180177@pilani.bit-pilani.ac.in (A. Agarwal); yash@pilani.bits-pilani.ac.in (Y. Sharma)
https://www.bits-pilani.ac.in/pilani/yash/profile (Y. Sharma)

Table 1
Training Data Distribution

	FAC	RLC	ARG	STA	PRE	Ratio	RPC
samples	2219	316	845	646	1468	3624	262

the evaluated macro F-scores, with a score of 0.442 in comparison to the 1st rank achieving an F-score of 0.468.

2. Related Work

Bhattacharya et al.[1] concluded in their paper that deep learning-based models perform much better at this task than using only handcrafted linguistic features, their approach made use of a heirarchial BiLSTM model. Attempts prior to [1] involved handcrafted feature extraction. For example, J. Savelka et al.[3] used CRF[4] and machine learning techniques on handcrafted features for automatically segmenting judgments into high functional and issue-specific parts. Nejadgholi et al.[5] proposed a skip-gram model[6] and a semi-supervised approach for search of legal facts in case documents using a classifier model from the fastText[7] library.

3. Dataset

The training dataset provided by AILA 2020 contained 50 annotated legal text documents, wherein every sentence was given a particular rhetorical role, as explained below:

1. **Facts (FAC)**: sentences that contain information related to the timeline of events that resulted in the case filing.
2. **Ruling by Lower Court(RLC)**: The cases mentioned in the dataset were given a preliminary ruling by the lower courts (Tribunal, High Court, etc.). These sentences correspond to the ruling/decision given by these lower courts.
3. **Argument(ARG)**: sentences that signify the arguments of the opposing parties
4. **Statute(STA)**: relevant statute cited
5. **Precedent(PRE)**: relevant precedent cited
6. **Ratio of the decision (Ratio)**: sentences that denote the rationale/reasoning given by the Supreme Court for the final judgment
7. **Ruling by Present Court(RLC)**: sentences that denote the final decision given by the Supreme Court for that case document

The train data contained 9380 training samples. The training data however, was not balanced in terms of number of samples per label, as shown in Table 1 so metric scores of labels with more samples are expected to be better than others. The test data contained 10 legal text documents without annotation, with a total of 1905 samples.

4. Proposed Technique

This method utilizes Transformers based models for the task of document classification. The proposed model uses a modified pretrained RoBERTa[8](a Robust and optimized BERT pretraining approach) encoder with an extra linear layer added to the pretrained RoBERTa base model, which was designed as an improvement of BERT[9] by providing advanced masked language modeling and significantly increasing the magnitude of training data.

4.1. Pretraining

The base model was pretrained in a self-supervised manner with the intention of a Masked language modeling (MLM) use case. It takes in a sequence of words as an input and masks 15% of them at random. This sentence is then fed into the model, which tries to predict the masked words. This pretraining approach lets the model learn the inner working/representation of the English language through a bidirectional approach towards learning the input sequence representation while training, this "language modeling" is beneficial for extracting features for other downstream tasks, which is text classification in our scenario. The datasets on which the base model was pretrained are as follows:

- BookCorpus
- Stories
- CC-News
- OpenWebText
- English Wikipedia

The above datasets are constituted from large amounts of unfiltered internet data. Therefore the train data is not entirely neutral in its modeling; thus, there are bound to be some biases in representing the English language.

4.2. Tokenization

A pretrained RoBERTa Tokenizer has been used to get the input ids and corresponding attention masks for each sentence. The tokenizer uses a byte variant of Byte-Pair Encoding (BPE)[10]. The tokenizer defines some special tokens to the input sequence. For example, a tokenized sentence always starts with the <s> token and is delimited by the </s> token.

15% of the tokenized sequence is masked. The masked tokens are then processed as follows:

- 80% of them are replaced with a <mask> token
- 10% are replaced with a random token different from the original one.
- The rest 10% are left as is.

As opposed to BERT, the model implements a dynamic masking approach that results in the masked token changing over the training epochs, making it robust for downstream tasks.

Table 2
Document wise metric scores

Document	macro precision	macro recall	macro F score
d1	0.639	0.621	**0.626**
d2	0.407	0.602	0.428
d3	0.662	0.726	**0.680**
d4	0.361	0.223	0.224
d5	0.580	0.601	**0.523**
d6	0.577	0.528	0.481
d7	0.463	0.545	0.490
d8	0.403	0.348	0.326
d9	0.278	0.257	0.230
d10	0.466	0.377	0.408

4.3. Fine-Tuning over a Legal Corpus

In the submitted run, the original pretrained base RoBERTa model in[8] was modified by adding a single linear layer on top for classification which was used as a sentence classifier. Training data was fed into the model in batches of 16, and the entire pretrained RoBERTa model and the additional untrained classification layer were fine-tuned for the particular downstream task of classifying the legal documents into one of the seven Rhetorical Roles. Adam Optimizer[11] was used while training with a learning rate of 2e-5 and epsilon value set to 1e-8. The said value was set to 42, and the model was fine-tuned for 4 epochs. The norms of the gradients were clipped to 1.0 to help prevent the explosive gradient problem.

Although RoBERTa trains on a dataset significantly larger in size than BERT, it reduces the computation times by employing techniques like distillation, pruning, etc. leading to a smaller network.

5. Results and Evaluation

The submitted model achieved an overall rank of 9 among all the submitted runs based on the F-score. The model was evaluated on the basis of classic classification metrics - macro averaged recall, precision and f-score. The metrics were calculated for each label for every document and then averaged over all the documents to get the overall results: Precision - 0.485, Recall - 0.483 and F1-score - 0.442. The document-wise metrics can be seen in Table 2. It can be observed that documents d1, d3 and d5 perform better than the rest of the documents, these documents have a lower number of data for which the train data was less for example, RLC, RPC and STA.

6. Conclusion and Future Work

The Transformers approach was chosen over a model that takes in data in a sequential manner (like LSTMs[12] or BiLSTM) to allow parallelism. The proposed model captures the English language context well since it was pretrained over a large corpus. However, the model is limited due to the specificity and the jargon encountered while fine-tuning it on the legal judgments dataset. The model seems to perform well in documents 1, 3, 5, which contained a lower number of test data for labels which had less train data, therefore given a well-balanced dataset to fine-tune on, the model could get better results than mentioned above. Future work on this approach can be to perform the pretraining of the language model on an extensive legal corpus. The pretrained tokenizer was also trained on generic English language data. Pretraining the tokenizer on a legal dataset would increase the quality of the embeddings being given to the transformer model.

References

[1] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, A. Wyner, Identification of rhetorical roles of sentences in indian legal judgments, 2019. arXiv:arXiv:1911.05405.

[2] 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: Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation, 2020.

[3] Savelka, K. D. Ashley, Segmenting u.s. court decisions into functional and issue specific parts, Frontiers in Artificial Intelligence and Applications 313 (2018) 111–120. URL: https://doi.org/10.3233/978-1-61499-935-5-111. doi:10.3233/978-1-61499-935-5-111.

[4] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: ICML, 2001.

[5] I. Nejadgholi, R. Bougueng, S. Witherspoon, A semi-supervised training method for semantic search of legal facts in canadian immigration cases, in: JURIX, 2017.

[6] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv:arXiv:1301.3781.

[7] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, ArXiv abs/1607.01759 (2017).

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019).

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019.

[10] T. Kida, S. Fukamachi, M. Takeda, A. Shinohara, T. Shinohara, S. Arikawa, Byte pair encoding: a text compression scheme that accelerates pattern matching, 1999.

[11] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014. arXiv:arXiv:1412.6980.

[12] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780. URL: https://doi.org/10.1162%2Fneco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735.