Classification of Sentences in Indian Legal Judgments

Vedant Parikh
DAIICT
India
201701076@daiict.ac.in

Prasenjit Majumder
DAIICT
India
p_mazumder@daiict.ac.in

ABSTRACT

Legal information is often represented in textual form (e.g., legal cases, contracts, bills). Hence, legal text processing is a growing area in NLP with various applications such as legal topic classification. Legal research is a process of finding relevant information from the judgments from various courts, law journals, etc. It is impossible for a person to manually go through each and every judgment to find the required information. Therefore, Automatic understanding of rhetoric roles of sentences in legal judgments is an important problem as it can be used in several downstream tasks like summarization of legal texts, legal search engine, information retrieval etc.

In this paper, Different approaches for Classification of legal sentences in Indian Legal Judgments are presented. We have used two different models for classification. We used Bidirectional Encoder Representation Transformer (BERT) Classification Model and Hierarchical Bi-LSTM plus Conditional Random Fields(CRF). In case of latter model we tried different pretrained embedding, namely, Sent2Vec, Universal Sentence Encoder, and BERT. The Data set consists of 50 human annotated Legal case Judgments from Supreme Court of India.

KEYWORDS

Semantic Segmentation, Rhetorical Roles, Legal Case Documents, Deep Learning, BiLSTM, BERT, Contextual Embeddings, Transfer Learning

1 Introduction

Classification of the sentence in their respective rhetorical roles requires understanding of semantic classes it belongs to such as, Facts, Statue, Argument, etc. This is can help in a variety of downstream tasks like semantic search, summarization, case law analysis, and so on. However there are two primary problems which makes this problem difficult to solve. Firstly, The structure of the legal judgments is ill - structured. The structure of the judgments depend on the legal domain of the cases. So, therefore we need to design domain specific algorithms and datasets. Secondly, the absence annotated Data set for sentence classification. This is due to the subjective nature of the rhetorical roles ,i.e, they may interleave with each other. Hence it sometimes becomes difficult even for human experts to understand the intricate differences between these roles. Another reason is hiring legal expertise is also expensive. As Supervised Learning require a large amount annotated data set. It is important to develop a high quality gold standard corpus for accurately capturing the rhetorical roles and also there is need to come up with more refined Deep learning Algorithm requiring less annotated Datasets such as unsupervised learning.

2 Dataset

Rhetorical Roles	Number of Sentences
Ratio of the decision (Ratio)	3624
Facts(FAC)	2219
Precedent (PRE)	1468
Argument (ARC)	844
Statue (STA)	646
Ruling by Lower Court (RLC)	315
Ruling by Present Court (RPC)	262

Table 1: Shows the 7 different Classes and Number of Sentences in respective Classes.

We consider legal judgments from the Supreme Court of India, from the website of Thomson Reuters Westlaw India.[1]Dataset consists of 50 documents from these 5 domains in proportion to their frequencies. Thus we have the following set of 50 documents from 5 domains – (i) Criminal – 16 documents (ii) Land and property – 10 documents (iii) Constitutional– 9 documents (iv) Labour and Industrial – 8 documents (v) Intellectual Property Rights – 7 documents. All experiments reported in this paper are performed on these 50 case documents. Dataset was annotated by the Law students. They have divided sentences into 7 classes shown in Table 1.

Dataset is not balanced. It can be inferred from Table1. The legal cases are contested by presenting the Arguments, Facts, Ratio of decision(Basically an argument) in front of a Judge. Therefore number of the sentences would be more as compared to Statue, Ruling of Lower Court and Ruling of Present Court. There is Huge variation in the length of the Sentences Fig 1.

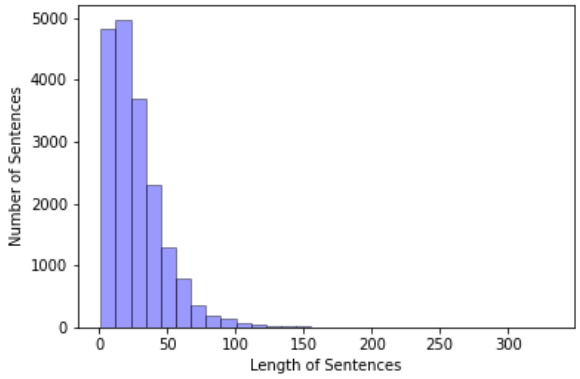

Fig 1: Showing the Variation in the length of Sentences in Legal Judgments.

3 Models

BERT: BERT is a language model based on Transformers pretrained on large corpora .For a new task, a task-specific layer is added on top of BERT and is trained jointly by fine-tuning on task-specific data. Its architecture is a multi-layer bidirectional Transformer encoder. It consists of 12 Encoder layers stack over one another, with 12 self Attention heads.We add a linear layer on top of BERT, with a sigmoid, softmax for Sentence Classification. BERT can process texts up to 512 tokens. So for long sentences we need to truncated. As we can see in Fig 1 that the length of sentences are mostly concentrated from Lengths 0 to 130. So, the sentences greater then 130 are removed. By doing so we are left with 9341 sentences. We have used **Bert-Base-Uncased** Pretrained Model for fine tuning. BERT has been pre-trained from large unlabeled text which we can use for downstream supervised task of sentence classification into their respective rhetorical roles.Therefore, BERT does not require large annotated Dataset. This approach can be used to solves the problem of not having large labeled dataset.

Hierarchical BiLSTM + CRF: BiLSTM model is used for Sentence classification into rhetorical roles and results from it is passed to Conditional Random Fields(CRF). The probability scores generated by the BiLSTM model do not take into account label dependencies, and thus can be regarded as simple emission scores. To enrich the model further, we deploy a CRF on top of the Hierarchical BiLSTM architecture.This requires us to feed the sequence of sentence embeddings to the BiLSTM, which returns a sequence of feature vectors. The BiLSTM model needs some initialization of the sentence embeddings, with which learning can start. We try two variations of sentence embeddings: (1) BERT Embeddigns by taking mean of last hidden state. (2) Universal Sentence Encoder.

We have used Hierarchical BiLSTM + CRF with pre-trained Sent2vec model as out Baseline results for comparing out results[1].

4 Evaluation Metrics

For a particular sentence, the label (rhetorical role) predicted by a model is considered to be correct, if it matches with the label assigned by the majority opinion of the human annotators. We use standard metrics for evaluating the performance of algorithms macro-averaged Precision, Recall and F-score. For macro-averaged metrics, we compute these metrics for each class separately, and then take their average (to prevent any bias towards the high-frequency classes).

5 Experiments, Results and Analysis

We now compare the performance of the models (stated in the previous section) on the set of 50 manually-annotated documents.

5.1 BERT Classification Model

Experiment 1:
Length of the Sentence : 64
K-Folds : 5
Number of Classes : 7
Epochs : 2

Experiment 2:
Length of the Sentence : 64
K-Folds : 5
Number of Classes : 7
Epochs : 4
As the number of epochs increased from 2 to 4, macro f1-score also increased. In this experiment **I got f1-score for Statue comparable with the baseline results.**

Experiment 3: Named Entities in the sentences were replaced by their respective class. We used spacey tools for finding the named entities in the sentences.

Length of the Sentence : 64
K-Folds : 1
Number of Classes : 7
Epochs : 4
The macro f1 score decreases. The replacement mostly effects the Ruling of Present Court and Ruling of Lower Court.

Experiment 4: As we see from the Fig 1 that the length of sentences vary from 0 to 385. And Max Sentences lie in the 0-130 range. I tried to increase the length of Sentence to 128.

Length of the Sentence : 128
K-Folds : 1
Number of Classes : 7
Epochs : 4
We can infer from the results that by taking longer sentences the macro f1-score increases. Maximum increase is seen in value of f1-score for Ruling in Lower Court and Ruling of Present Courts.

Experiment 5: As we can see from Table 1 that the dataset is Imbalanced. So I tried to make the dataset more balanced by taking only 4 classes having similar nos. of Sentences, i.e, around thousand. The 4 Classes taken are Facts, Ratio of Decision, Precedent, Argument.

Length of the Sentence : 64
K-Folds : 1
Number of Classes : 4
Epochs : 4
As the Dataset becomes more balanced, the macro-f1 score for only the four classes increases to 0.62. Therefore we require more annotated dataset for other 3 classes.

Experiment 6: In Experiments 1 to 5 we have used sentence level splitting. In Experiment 6, We have tried for Document Level Splitting.

Length of the Sentence : 64
K-Folds : 1
Number of Classes : 7
Epochs : 4
The f1-score for argument in the experiment is comparable with the baseline results.

Experiment 7: In Experiment 6, We have tried for Document Level Splitting. We increase the sentence length from 64 to 220.

Length of the Sentence : 220
K-Folds : 1
Number of Classes : 7
Epochs : 4
In this experiment as the length of sentences is increased, the macro f1-score is drastically decreased.

Experiment 8: I Increased the number of epochs, to see if the the accuracy is increasing or not.

Length of the Sentence : 220
K-Folds : 5
Number of Classes : 7
Epochs : 10
As we increase the epochs the macro f1-score increases, but over fitting of model takes place. As Average training loss is 0.02 but Average Validation loss is 2.1.
(Results shown in Table 2 on the Last Page)

5.2 Hierarchical BiLSTM + CRF

For this model we require pre-trained embeddings of the sentences. So we have used 2 different Sentence Embeddings for this model.

5.2.1 BERT Embeddings

The BERT embeddings are the contextual embeddings trained by fine tuning the BERT. It takes both Right to Left and Left to Right dependencies of words into account. It also uses self Attention which also helps in long distance relativity between the tokens.

5.2.2 Universal Sentence Encoder

The Universal Sentence Encoder encodes text into high dimensional vectors that can be used for text classification, semantic similarity, clustering, and other natural language tasks. It is trained by using Transformer encoder. Here we have used pre-trained embeddings of 512 dimensions. Hyperparameters used in different embeddings. K-Folds : 5, Number of Classes : 7, Epochs : 300, Learning Rate: 0.01, Batch Size: 32.

The difference between the two embeddings used is that the BERT embeddings are contextual(takes surrounding words into account) while Uniform Sentence Encoder is Non-contextual embeddings. Therefore, we can see from the results that BiLSTM with contextual embeddings better than non-contextual embeddings. (Results shown on the last page in Table 3).

So, we can conclude from the results that class 'Ruling in the Lower court' has performed worst with highest f1-score across all the models. So more balanced annotated dataset must be created. For future works we can try BERT-LARGE-UNCASED model for fine tuning. We can pre-train BERT using a large corpus of Legal judgments. We also found that BERT Classification Model and Hierarchical BiLSTM with fine tuned BERT Embeddings.

ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to Prof. Prasenjit Majumder and Teaching Assistant Apurv Sir for their able guidance and support in completing my Project.

REFERENCES

[1] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, AdamWyner. "Identification of Rhetorical Roles of Sentences in Indian Legal Judgments" in arXiv:1911.05405.

[2] Ilias Chalkidis, Ion Androutsopoulos, Nikolaos Aletras."Neural Legal Judgment Prediction in English"

[3] Lisa Andreevna Chalaguine,Claudia Schulz. "Assessing Convincingness of Arguments in Online Debates with Limited Number of Features"

[4] Filip Boltuzic and Jan Snajder. "Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity"

[5] Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and R'emi Louf and Morgan Funtowicz and Jamie Brew "HuggingFace's Transformers: State-of-the-art Natural Language Processing".ArXiv 2019 abs/1910.03771

[6] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina."BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" in arXiv preprint arXiv:1810.04805

[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong. "Universal Sentence Encoder" in arxiv.org/abs/1803.11175.

Experiments	FAC	ARG	Ratio	STA	PRE	RPC	RLC	Macro Average F1 score
Baseline Results	0.8388	0.5924	0.92	0.7218	0.8538	0.9002	0.812	0.8208
1	0.6655	0.5213	0.6703	0.6904	0.4537	0.6074	0.029	0.5196
2	0.6535	0.5237	0.64798	**0.7465**	0.55802	0.5348	0.2168	0.5544
3	0.66	0.46	0.63	0.72	0.48	0.36	0.14	0.49
4	0.69	0.54	0.66	0.67	0.55	0.57	0.36	0.58
5	0.72	0.51	0.73	-	0.53	-	-	0.62
6	0.6467	**0.5804**	0.5706	0.6694	0.5381	0.5333	0.19	0.5334
7	0.59	0.3396	0.5745	0.42	0.29	0.58	0.17	0.42
8	0.674	0.5264	0.6906	0.6643	0.6078	0.63	0.242	0.5764

Embeddings Used	FAC	ARG	Ratio	STA	PRE	RPC	RLC	Macro Average F1 Score
BERT	0.744	0.5312	0.6772	0.701	0.571	0.704	0.201	0.5899
Universal Sentence Encoder	0.71	0.482	0.6726	0.653	0.598	0.6996	0.233	0.485

Table 3: Results shown for Hierarchical BiLSTM for 2 pre-trained embeddings stated in prior section.

Table 2: Results shown for BERT Classification Modal for 7 Experiments stated in prior section.