Rhetorical Role Labelling for Legal Judgements Using ROBERTA

Soumayan Bandhu Majumder[a], Dipankar Das[a]

[a] *Jadavpur University, 188, Raja S.C. Mallick Rd, Kolkata, India*

Abstract

In the present attempt we build a Roberta based model for shared task 2 (Rhetorical Role Labelling for Legal Judgements) in AILA 20. We use Roberta model to get the text embedding. Output of the Roberta is passed through the bidirectional LSTM of 256 units and 128 units, after that passed that output through a dense layer and global_max_pooling_1D layer and finally a softmax layer of 7 activation unit. We use batch size of 16 and max length of 120. We submit 3 runs where 2nd, 3rd and 1st run submissions were scored (Macro-F-score) 0.468, 0.457 and 0.452 respectively. These three systems ranked 1st, 2nd and 4th respectively.

Keywords 1

Legal document, Roberta, Rhetorical role

1. Introduction

We participated in AILA (Artificial Intelligence for Legal Assistance) shared task 2 of FIRE (Forum for Information Retrieval Evaluation) conference. The task is to semantically segment a legal case document. More formally, it is a sentence classification task, where each sentence has to be assigned one of the 7 predefined labels or rhetorical roles. This 7 predefined labels are Facts, Ruling by lower court, Argument, Statue, Precedent, Ratio of the decision and Ruling by present court. We try different types of models but Roberta based model gives best output. We submit 3 different models based upon Roberta with 3 different epochs. We trained for 13 epochs for our 1st system (ranked 4, Macro-F score 0452), 2nd system trained for 15 epochs (ranked 1, Macro-F score 0.468) and 3rd system trained for 19 epochs (ranked 2, Macro-F score 0.457). In our system we first use Roberta then passed that output through bilstm, dense layer, global_max_pooling_1D layer and finally through the softmax layer to get the output.

The rest of the paper is organized as follows. In Section 2 Related work on this particular topic is discussed. Whereas Section 3 briefly shows the insights of the datasets. Section 4 describes the method we used to classify each sentences of legal documents and also describes our models. Section 5 is dedicated to experiments and results. Finally, in Section 6, we present the conclusions and briefly discuss about future work.

2. Related Study

There are some previous work already done on this area. Saravanan et.al. [1] worked upon rhetorical role identification in rent control, income tax, sales tax domain. Here they have used Conditional Random Field (CRF) upon label transition features. Savelka et.al. [2] also worked upon in rhetorical role identification and used CRF upon parts of speech tags. They here worked upon Cyber crime domain. Bhattacharya et.al. [3] worked upon legal domain for rhetorical role identification and used LSTM with CRF to detect each of seven labels.

Forum for Information Retrieval Evaluation 2020, December 16-20, 2020, Hyderabad, India
EMAIL: soumayanmajumder@gmail.com (A. 1); dipankar.dipnil2005@gmail.com (A. 2)

3. Dataset

Legal case documents follow a common thematic structure with implicit sections like Facts of the Case, Issues being discussed, Arguments given by the parties, etc. These sections are popularly termed as "rhetorical roles".

So here we have to classify each sentence of a document to one of the below mentioned seven classes. Here we briefly explain what those seven classes actually mean for better understanding.

1. **Facts**: Sentences that denote the chronology of events that led to filing the case.
2. **Ruling by Lower Court**: Since we will be providing Indian Supreme Court cases, these cases 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**: Sentences that denote the arguments of the contending parties.
4. **Statute**: Relevant statute cited.
5. **Precedent**: Relevant precedent cited.
6. **Ratio of the decision**: Sentences that denote the rationale/reasoning given by the Supreme Court for the final judgement.
7. **Ruling by Present Court**: Sentences that denote the final decision given by the Supreme Court for that case document.

We are given 50 legal documents for training purpose and 10 legal documents for testing. Number of sentences present in these documents are different. Each sentence is labelled with one of the seven classes. These classes are not balanced. Below we give a snapshot of our dataset that is provided us for training.

	text	target
0	The present appeal arises out of the judgment ...	Facts
1	302 of the Indian Penal Code (for short the 't...	Facts
2	Facts giving rise to the present appeal may be...	Ratio of the decision
3	On 19.04.1988 between 8.30 p.m. and 8.45 p.m.,...	Facts
4	It is also the case of the prosecution that as...	Facts
5	P.S. Joaquim Dias (PW-21) who was attached to ...	Facts
6	He was informed that the deceased had died as ...	Facts
7	On receipt of the aforesaid message, PW-21 alo...	Facts
8	They reached the scene of offence at about 11....	Facts
9	During the survey made at the place of occurre...	Facts

Snapshot of our training dataset

4. Methods

Pre-processing:- We are given labelled dataset. Labels of these dataset are categorical variables so first we have to convert these variables to one hot encoded variables. Here we have done nominal encoding (one hot encoding) instead of ordinal encoding because here order does not matter.

Model:- We have tried different types of models for this shared task. Mainly those models are Random forest, Universal Sentence Encoder, BERT, ROBERTA. Among these models ROBERTA gives the best result.

In Random forest based model we used around 500 trees which are of full depth with gini index.

In Universal Sentence encoder based model we have used U.S.E. large version 5, first we get those embeddings and then passed through the softmax layer of 7 activation unit to get the desired output.

In BERT based model we used bert_large_uncased. We passed the output of the bert through the 256 unit neural network with relu activation unit then apply the dropout of 0.4 and passed it through the 128 unit neural network with relu activation unit and again apply dropout of 0.4 and finally passed it through the softmax layer of 7 activation unit to get the desired output.
In Roberta based model to predict labels the output of Roberta is send to the bidirectional LSTM of 256 units and 128 units respectively, after that passed that output through a dense layer and global_max_pooling_1D layer and finally a softmax layer of 7 activation unit. We have used loss function as categorical crossentropy and Adam as optimizer. We use batch size of 16 and max length of 120. We submit 3 runs where 1st run submission runs for 13 epoch, 2nd run submission runs for 15 epoch and 3rd run submission runs for 19 epoch.

5. Experiments & Results

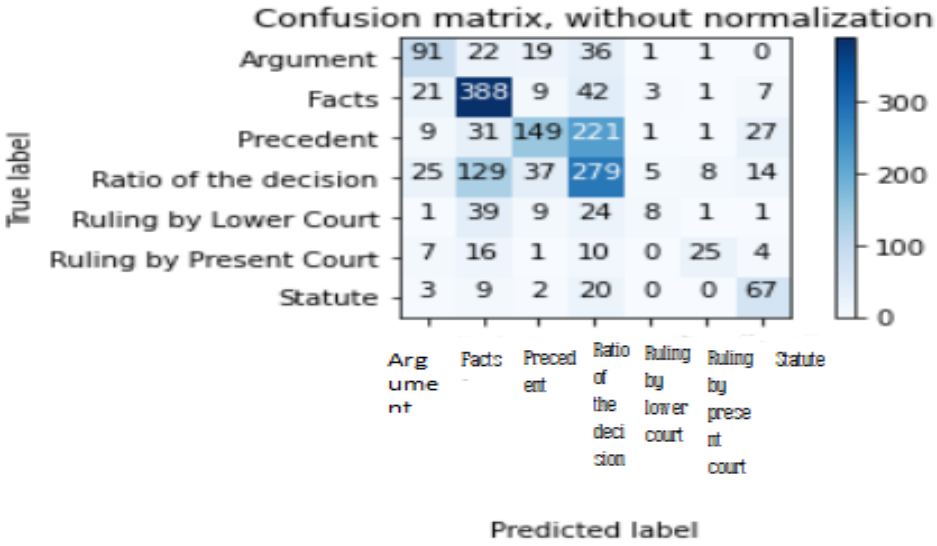
Here we are going to see performance of our four different models (Random forest, Universal Sentence Encoder, BERT, ROBERTA) upon validation dataset, which we taken from our training dataset. From the below mentioned table we can see that among these models ROBERTA based model performed best.

Models	Metrics	Argument	Facts	Precedent	Ratio of the decision	Ruling by Lower Court	Ruling by Present Court	Statute
Random Forest	precision	0.103	0.74	0.6	0.52	0	1	0.75
	recall	0.94	0.036	0.006	0.25	0	0.174	0.06
	F score	0.186	0.068	0.013	0.336	0	0.297	0.11
Universal Sentence Encoder	precision	0.8	0.55	0.73	0.42	0	0.94	0.511
	recall	0.165	0.82	0.30	0.67	0	0.24	0.45
	F score	0.27	0.66	0.42	0.51	0	0.38	0.48
BERT	precision	0.87	0.63	0.76	0.44	0	0.76	0.74
	recall	0.54	0.84	0.36	0.71	0	0.444	0.367
	F score	0.67	0.722	0.49	0.54	0	0.56	0.49
ROBERTA	precision	0.58	0.612	0.66	0.44	0.444	0.68	0.56
	recall	0.54	0.82	0.34	0.56	0.10	0.40	0.66
	F score	0.56	0.70	0.45	0.50	0.16	0.50	0.61

Different precision, recall, F-score produced by different classes and different models

From the below mentioned confusion matrix (of ROBERTA model) we can see that from our validation dataset (10 documents of training set) which classes are predicted correctly and which are predicted wrong. We can see that Argument class predicted correctly 91 times, Facts class predicted correctly 388 times, Precedent class predicted correctly 149 times but precedent predicted as Ratio of the decision 221 times that is why we are getting recall value of 0.34, Ratio of the decision is correctly predicted 279 times but wrongly predicted as Facts 129 times, Ruling by Lower court predicted correctly for just 8 times and this creates the most problem for our models. This Ruling by Lower court predicted as Facts 39 times and Ratio of the decision 24 times so its recall value decreases very much and become 0.10 only. Ruling by present court predicted correctly 25 times and Statute predicted correctly 67 times.
For evaluation of the model Macro F-score is considered instead of accuracy because here given classes are not balanced. Here task organizers first calculate Recall, Precision and F-score for each category of

labels within each document. Then the score for each document in a run were computed by averaging the scores for all seven categories in that document. Finally, the overall scores for a run are computed by averaging the scores for each document.



Team_runid	Macro precision	Macro recall	Macro F-score	Accuracy
ju_nlp_1	0.504	0.483	0.452	0.588
ju_nlp_2	**0.506**	**0.501**	**0.468**	**0.588**
ju_nlp_3	0.519	0.479	0.457	0.57

Final result given by the task organizers

Our system ju_nlp_2 ranked 1 in this competition with Macro F-score of 0.468. We submit our Roberta model with 13 epochs (ju_nlp_1), 15 epochs (ju_nlp_2) and 19 epochs (ju_nlp_3) respectively.

6. Conclusion

From the above we can see that we try different types of models to get desired output but we face main problem to detect the class Ruling by the lower court because in our validation dataset (comprised of 10 documents) we can only predicted it 8 times correctly and precision and recall value of 0.444 and 0.10 respectively. In future we should try some other things to detect this particular class correctly. Performance of this model depends upon how the test data reflects the real world dataset.

7. References

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

[2] J. Savelka and K. D. Ashley, "Segmenting U.S. court decisions into functional and issue specific parts," in Proc. JURIX, 2018.

[3] Bhattacharya, Paheli & Paul, Shounak & Ghosh, Kripabandhu & Ghosh, Saptarshi & Wyner, Adam. (2019). Identification of Rhetorical Roles of Sentences in Indian Legal Judgments.

[4] 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 Proc. of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020.