FIRE 2020 AILA Track: Artificial Intelligence for Legal Assistance

Paheli Bhattacharya
Indian Institute of Technology
Kharagpur, India
paheli.cse.iitkgp@gmail.com

Parth Mehta
Parmonic AI
parth.mehta126@gmail.com

Kripabandhu Ghosh
Indian Institute of Science Education
and Research (IISER), Kolkata, India
kripa.ghosh@gmail.com

Saptarshi Ghosh
Indian Institute of Technology
Kharagpur, India
saptarshi@cse.iitkgp.ac.in

Arindam Pal
Data61, CSIRO and Cyber Security
CRC, Sydney, NSW, Australia
arindamp@gmail.com

Arnab Bhattacharya
Indian Institute of Technology
Kanpur, India
arnabb@iitk.ac.in

Prasenjit Majumder
DA-IICT Gandhinagar, India
prasenjit.majumder@gmail.com

ABSTRACT

The FIRE 2020 AILA track aimed at developing datasets and frameworks for the following two tasks: (i) Precedent and Statute Retrieval, where the task was to identify relevant prior cases and statutes (written laws) given a factual scenario, and (ii) Rhetorical Role Labelling for legal judgements, where given a case document, sentences were to be classified into 7 rhetorical roles – Fact, Ruling by Lower Court, Argument, Precedent, Statute, Ratio of the decision and Ruling by Present Court. For both the tasks, we used publicly available Indian Supreme Court case documents.

KEYWORDS

Legal data analytics, Prior case retrieval, Statute retrieval, Legal facts, Rhetorical Role labelling, Semantic Segmentation, Legal Information Retrieval

ACM Reference Format:
Paheli Bhattacharya, Parth Mehta, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2020. FIRE 2020 AILA Track: Artificial Intelligence for Legal Assistance. In *Forum for Information Retrieval Evaluation (FIRE '20), December 16–20, 2020, Hyderabad, India.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3441501.3441510

1 INTRODUCTION

In this paper, we describe in brief the 'Artificial Intelligence for Legal Assistance' (AILA) track at the Annual Conference of the Forum for Information Retrieval Evaluation (FIRE 2020 – http://fire.irsi.res.in/fire/2020/home). The track focused on two tasks: (1) Identifying relevant prior cases and statutes, given facts of a legal situation (this task is the same as that of AILA 2019 [2], with an extended dataset), and (2) Rhetorical Role Labeling for legal judgements (this task is newly introduced in AILA 2020). The objective is to semantically segment a long and unstructured case document.

Both the above tasks consider Indian legal documents, i.e., Indian statutes and prior cases decided by Indian courts of Law (the datasets are detailed in the subsequent sections). Note that Indian legal case judgments are not written in a structured way, and there are no section titles either. Hence it becomes a challenging task to identify specific portions such as the facts of the case, ruling of the court, etc. [3].

We briefly describe each of these tasks in the subsequent sections. More detailed explanation of the problem definition and description of the approaches of the participating teams can be found in [4].

FIRE '20, December 16–20, 2020, Hyderabad, India

ACM ISBN 978-1-4503-8978-5/20/12…$15.00
https://doi.org/10.1145/3441501.3441510

2 TASK 1: PRECEDENT AND STATUTE RETRIEVAL

Given a set of queries, each of which describes (in natural English language) a situation that had led to filing a case in an Indian court of law, the task was to find :

(1) Relevant Precedents from a pool of 3,257 case documents that were judged in the Supreme Court of India. For each query, the task was to retrieve the most similar / relevant case documents with respect to the situation in the given query.
(2) Relevant Statutes or written laws/Acts that are relevant to the legal situation. We identified a set of 197 statutes (Sections of Acts) from Indian law, that are relevant to some of the queries stated above. We provided the participants with the title and description of these statutes. For each query, the task is to identify the most relevant statutes (from among the 197 statutes).

This task was a continuation of the AILA 2019 track. Details about the dataset construction can be found in [2].

2.1 Training and Test Data

The tasks of identifying relevant Precedents and Statute can be modeled either as unsupervised retrieval tasks (where one searches for relevant statues/prior cases) or as a supervised classification task

Team name	Run ID	P@10	MAP	BPREF	recip_rank	Method Summary
UB	UB-3	0.08	0.1573	0.1128	0.238	Query expansion using Terrier 4.2 KL divergence model
fs_hu	fs_hu_task1a	0.10	0.1351	0.0885	0.2041	Language model and Dirichlet Smothing
double_liu_2020	double_liu_2020_1	0.07	0.1306	0.0737	0.1963	Extract search keywords using IDF, BM25

Table 1: Results of Task 1: Precedent retrieval for queries. All measures averaged over 10 test queries. Rows are sorted in decreasing order of MAP score (primary measure).

Team name	Run ID	P @ 10	MAP	BPREF	recip_rank	Method Summary
scnu	scnu_1	0.18	0.3851	0.3054	0.5615	BERT for training
SSNCSE_NLP	task_1b_2	0.07	0.3423	0.136	0.3423	TF-IDF
IMS_UNIPD	tfidf_stem	0.17	0.3383	0.279	0.5349	TF IDF weighting on stemmed words

Table 2: Results of Task 1: Statute retrieval for queries. Measures averaged over 10 test queries. Rows are sorted in decreasing order of MAP score (primary measure).

(where one tries to predict for each statute/prior case whether it is relevant to the given query). To facilitate considering the tasks as supervised learning tasks, we provided the dataset of AILA 2019 [2] as the training set. The training data had 50 queries. For each of these queries, the relevant prior cases and statutes were labelled. Note that, we increased the pool of prior case documents from 2, 914 (in AILA 2019) to 3, 257 (in the present AILA 2020 track). The pool of statutes were the same in both tracks, i.e., 197.

For the test data, we created 10 more queries, along with their relevant prior cases and statutes. The dataset creation process was same as that in AILA 2019 [2].

2.2 Evaluation Methodology

For the relevant prior case/statute retrieval task, a submitted method generated a ranked list of documents (prior cases/statutes) relevant to each query. One set of ranked lists (one for each query) generated by a certain method is called a 'run'. We evaluate the submitted runs, based on their performance over all 10 test queries.

The evaluation procedure was similar to AILA 2019 [2] – we used Mean Average Precision (MAP), Precision@10 (P10), BPREF and Reciprocal rank (recip_rank) as the evaluation metrics. The *trec_eval* tool [1] tool was used for computing the metrics stated above. We choose MAP as the primary measure since it incorporates both Precision and Recall.

2.3 Best Performing Approaches

For the first task of retrieving relevant prior/ precedent cases, we received a total of 26 runs from 10 participating teams. For the second task of retrieving relevant statutes , we received a total of 27 runs from 12 participating teams. We provide a brief overview below, of the methodologies used by the teams with the best performing runs. The results of top three best performing runs are in Tables 1 and 2. More detailed descriptions of approaches used by all the teams and their evaluation scores can be found in [4].

The top three runs for Task 1 (Identifying relevant prior cases) were from teams UB [8], fs_hu [9] and double_liu_2020 [10]. For their best performing run, team UB used query expansion using Terrier 4.2 *KL divergence* model. The approach from team fs_hu (which ranked second in the overall rankings) used Language model and Dirichlet Smoothing. The third ranking team double_liu_2020 extracted the top 50% of top IDF as the search key for each query and use BM25 as the search score.

For the second task of retrieving relevant statutes, the top three runs were from teams scnu[2], SSNCSE_NLP [1] and IMS_UNIPD [6]. The results are in Table 2. The top performing team scnu used a supervised training mechanism using BERT. Note that, we had provided AILA 2019 dataset comprising of 50 queries with the relevant prior cases and statutes labelled. This was the only team which utilized this data for training a machine learning model. The next best performing teams SSNCSE_NLP and IMS_UNIPD used TF-IDF based weighting. More detailed descriptions of approaches used by all the teams and their evaluation scores can be found in [4].

3 TASK 2 : RHETORICAL ROLE LABELING FOR LEGAL JUDGEMENTS

Legal case documents are usually long and unstructured, with little or no section headings, making them difficult to read. It becomes tedious for a reader to understand where the facts of the case is written, where are the arguments by the contending parties mentioned and so on. Therefore, the task of semantic/thematic segmentation also known as rhetorical role labelling of sentences, becomes an important task. It not only enhances the readability of the document but also has applications in several downstream tasks like summarization, case law analysis, semantic search and so on. We consider the following seven (07) rhetorical labels/semantic segments [5]:

- Facts : legal situation that led to filing the case
- Ruling by Lower Court : since we consider documents from the Supreme Court of India, there was some preliminary ruling given at the lower courts e.g.. High Court, Tribunal etc.
- Argument : arguments made by the contending parties
- Precedents : citation to relevant prior cases
- Statutes : citation to relevant statutes
- Ratio of the decision : reasoning behind the final judgement
- Ruling by Present Court : final judgement given by the Supreme Court of India

[1] https://trec.nist.gov/trec_eval/

[2] The team scnu decided not to submit their working notes.

Team	Run ID	P	R	F	Acc	Method Summary
ju_nlp	ju_nlp_2	0.506	0.501	0.468	0.588	ROBERTA
heu_gjm	heu_gjm_1	0.541	0.472	0.457	0.603	BERT + Logistic Regression
double_liu	double_liu_3	0.472	0.486	0.444	0.619	BERT

Table 3: Results of Task 2: Rhetorical Role Labeling for Legal Judgements. Measures averaged over 10 test documents comprising of 1,905 sentences. Rows are sorted in decreasing order of FScore (primary measure).

The task here is to label each sentence of a case document with one of the above rhetorical roles.

3.1 Training and Test Data

As the training set, we provided a set of 50 Indian Supreme Court case documents where each sentence was labelled with one of the above 7 rhetorical roles. There were 9,308 sentences in total. This dataset was made available by our prior work [5].

As the test set, we curated a set of 10 additional case documents. We randomly selected 2 documents from each of the 5 law domains mentioned in [5]. These documents were then given to a law expert for annotating every sentence with one of the rhetorical labels. There are a total of 1,905 sentences in the test set.

3.2 Evaluation Methodology

We used the standard Recall, Precision and F1-Scores. The documents have a considerable variation in their size. Moreover, even within a document, there is a class imbalance among the 7 categories / rhetorical roles. Hence we use macro-averaging at both document-level and category-level. The scores were calculated as below:

(1) Recall, Precision and F-score were computed for each category of labels within each document.
(2) The score for each document in a run were computed by averaging the scores for all seven categories in that document.
(3) Finally, the overall scores for a run are computed by averaging the scores for each document.

We additionally report the overall Accuracy for each submitted run. Accuracy is micro-averaged across documents and classes, i.e., it is measured as the fraction of sentences correctly classified out of the 1,905 test sentences.

3.3 Best Performing Approaches

For the Rhetorical role labelling task we received a total of 21 runs from 9 participating teams. The results of some of the top-performing runs are shown in Table 3. We provide a brief overview below, of the methodologies used by the teams with the best performing runs.

The top-performing runs were all transformer-based models or some modifications of them. Three out of the top 5 best performing runs (ranked 1, 2 and 4) were submitted by the team JU [11]. These runs actually use the same technique – ROBERTA transformer – but use different number of epochs. In order to report more diverse techniques, we report runs from three distinct top performing teams in Table 3. The other two top-performing teams are heu_gjm[7] and double_liu [10]. While heu_gjm used BERT-based features combined with Logistic Regression, the team double_liu again used BERT to obtain the results. The approach by double_liu was also the best performing approach in terms of Accuracy. More detailed descriptions of approaches used by all the teams and their evaluation scores can be found in [4].

4 CONCLUSION

The FIRE AILA track is meant for creating benchmark collections for Legal-AI tasks such as identification of relevant statutes and prior cases, rhetorical segmentation of case documents, and so on. As evident from the result tables, we find that these tasks are indeed challenging, having lot of scope of improvement. We plan to continue the track in future, to strive towards developing better methods for these and other AI-Legal tasks.

Acknowledgements: The track organizers thank all the participants for their interest in this track. We also thank the FIRE 2020 organizers for their support in organizing the track. The research is partially supported by SERB, Government of India, through a project titled "NYAYA: A Legal Assistance System for Legal Experts and the Common Man in India". P. Bhattacharya is supported by a PhD Fellowship from Tata Consultancy Services.

REFERENCES

[1] Nitin Nikamanth Appiah Balaji, B. Bharathi, and J. Bhuvana. 2020. Legal Information Retrieval and Rhetorical Role Labelling for Legal Judgements. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*
[2] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Overview of the FIRE 2019 AILA track: Artificial Intelligence for Legal Assistance. In *Proceedings of FIRE 2019 - Forum for Information Retrieval Evaluation.*
[3] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. In *Proc. European Conference on Information Retrieval (ECIR).*
[4] Paheli Bhattacharya, Parth Mehta, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2020. Overview of the FIRE 2020 AILA track: Artificial Intelligence for Legal Assistance. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*
[5] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of Rhetorical Roles of Sentences in Indian Legal Judgments. In *Proc. International Conference on Legal Knowledge and Information Systems (JURIX).*
[6] Giorgio Maria Di Nunzio. 2020. A Study on Lemma vs Stem for Legal Information Retrieval Using R Tidyverse. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*
[7] Jamming Gao, Hui Ning, Zhongyuan Han, Leilei Kong, and Haoliang Qi. 2020. Legal text classification model based on text statistical features and deep semantic features. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*
[8] Tebo Leburu-Dingalo, Nkwebi Peace Motlogelwa, Edwin Thuma, and Monkgogi Modungo. 2020. UB at FIRE 2020 Precedent and Statute Retrieval. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*
[9] Zhiran Li and Leilei Kong. 2020. Language Model-based Approaches for Legal Assistance. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*
[10] Liang Liu, Lexiao Liu, and Zhongyuan Han. 2020. Query Revaluation Method For Legal Information Retrieval. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*
[11] Soumayan Bandhu Majumder and Dipankar Das. 2020. Rhetorical Role Labelling for Legal Judgements Using ROBERTA. In *Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation.*