AILA 2021: Shared task on Artificial Intelligence for Legal Assistance

Vedant Parikh
DA-IICT, Gandhinagar and Amazon
India
vedant.parikh.6299@gmail.com

Upal Bhattacharya
Indian Institute of Science Education and Research
Kolkata, India
upal.bhattacharya@gmail.com

Parth Mehta
Parmonic AI
USA
parth.mehta126@gmail.com

Ayan Bandyopadhyay
TCG Crest
Kolkata, India
bandyopadhyay.ayan@gmail.com

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

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

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

Arindam Pal
Data61, CSIRO and University of New South Wales
Sydney, Australia
arindamp@gmail.com

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

Prasenjit Majumder
DA-IICT, Gandhinagar and TCG Crest, Kolkata
India
prasenjit.majumder@gmail.com

ABSTRACT

AILA 2021 was the third edition of the Shared task on Artificial Intelligence for Legal Assistance, that was organized with the FIRE 2021 conference. This year two tasks were offered. While the rhetorical role labelling task was continued from last year, a new Legal Judgement Summarization task was introduced in the current edition. In the Rhetorical Role Labelling task, given a case document, the sentences in the document 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. The legal judgement summarization task consisted of two subtasks. Subtask (a) was a binary classification task that required participants to identify 'summary-worthy' sentences in a court judgement. For subtask (b) participants had to automatically generate a summary from a given court judgement. Datasets for both tasks were created by annotating publicly available judgments from the Supreme court of India.

FIRE 2021, December 13–17, 2021, Virtual Event, India

ACM ISBN 978-1-4503-9596-0/21/12.
https://doi.org/10.1145/3503162.3506571

KEYWORDS

Legal data analytics, Rhetorical role labelling, Semantic segmentation, Legal document summarization, Headnote generation

ACM Reference Format:
Vedant Parikh, Upal Bhattacharya, Parth Mehta, Ayan Bandyopadhyay, Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2021. AILA 2021: Shared task on Artificial Intelligence for Legal Assistance. In *Forum for Information Retrieval Evaluation (FIRE 2021), December 13–17, 2021, Virtual Event, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3503162.3506571

1 INTRODUCTION

The series of shared tasks on Artificial Intelligence for Legal Assistance (AILA) are focused on relevant problems in Legal NLP and Text mining. AILA 2021 builds on previous tasks related to Legal NLP offered at the Forum for Information Retrieval Evaluation [1, 2, 8]. Previous tracks have focused on Legal IR, statute and precedent retrieval, and rhetorical role labelling. In this overview paper, we briefly describe the 3rd edition of AILA organized with the 12th annual conference of the Forum for Information Retrieval Evaluation (FIRE 2021)[1].

This year the focus was on two tasks: (1) Rhetorical Role Labeling for legal judgements (this task is continued from AILA 2020). The objective is to semantically segment a long and unstructured case document for better comprehension of the document as well as to aid downstream tasks. (2) Legal Document Summarization. This

[1] http://fire.irsi.res.in/fire/2021/home

task was newly introduced and focuses on automatic summarization of case documents.

For both tasks we use publicly available judgements delivered by the Supreme Court of India, with further annotations as described in subsequent sections. We briefly describe each of these tasks and the best performing submissions here.

2 TASK 1 - RHETORICAL ROLE LABELLING

Previous works by the track organizers [3] and a previous edition of AILA [2] make a strong case for rhetorical role labelling task. In a nutshell, a document annotated with rhetorical roles allows a reader to better understand what type of information is present where in the document. Such labeling 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. Following seven rhetorical role categories were used for this task:

- 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).
- 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

2.1 Training and Test Data

The training set consisted of 60 Indian Supreme Court case documents where each sentence was labelled with one of the above 7 rhetorical roles. The training set was partly made available by our prior work [3] and partly by the previous edition of AILA [2]. More specifically, the train and test datasets from the AILA 2020 rhetorical role labelling task were jointly used as the training set this year. Further, a validation set consisting of 10 documents was released to allow participants to fine tune their models. The test set consisted of a curated set of 10 additional documents. The strategy for selecting these documents was same as in previous edition.

2.2 Evaluation Methodology

The evaluation methodology was same as that in AILA 2020 [2]. Standard metrics of Recall, Precision and F1-Scores were used to rank the systems. Since there is a class imbalance among the 7 categories / rhetorical roles, we use macro-averaging at category-level. The scores were calculated as below:

(1) Recall, Precision and F-score were computed for each category of labels across all documents.
(2) The overall scores for a run are computed by averaging the scores across all categories.

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 total test sentences.

2.3 Best Performing Approaches

For task 1 we received 26 runs across 11 teams. Like last year, all the top performing runs used transformer-based approaches. All three runs from team Rustic [4] were ranked top-3 in the ranked list. However, here we list just the top performing run from this team and compare it with the next best runs from teams Minitrue[2] and Arguably [7] in Table 1. Both Rustic and Minitrue used the domain-specific pretrained model legal-BERT as the base model. Rustic further experimented with structural embeddings in addition to legal-BERT, while Minitrue used a simple neural inference network in addition to legal-BERT. Arguably used a different transformer-based model Ernie 2.0 for this task.

3 TASK 2 - LEGAL DOCUMENT SUMMARIZATION

Given a court judgement, not all parts of it are equally important from a laywer's perspective. For instance, often the facts and rationale of the judgement are given more importance while creating a headnote compared to an argument. In task-2a we aim to replicate this behaviour. Given a judgement, the task is to identify sentences which are "summary worthy", i.e. they have at least some information which should be included in the summary. This task can be seen as a binary sentence classification task.

Task-2b is an extension of task 2a, where the participants are required to automatically generate a summary, either extractive or abstractive, for a given judgement. In the simplest case the extractive summary could be formed by reordering the sentences identified as important in task 2a. The alternate to this would be to compress/re-write sentences from task-2a or to use generative models for creating truly abstractive summaries. Since the court judgements, and as a result the corresponding headnotes, can vary substantially in length we do not use a fixed length summary. Instead, for each judgement in the test dataset, the target summary length in number of words is provided.

3.1 Training and Test Data

The dataset for this task consists of judgements delivered by the Supreme Court of India alongside the headnotes, which are handwritten summaries of the judgements. Legal documents often include complex sentences and large number of acronyms, which makes the task of sentence tokenization non-trivial. Most common tokenizers that are readily available, such as nltk[3] or Spacy[4], usually do not provide accurate results for legal documents, and this inaccuracy often affects the downstream tasks. To mitigate this issue, we provide pre-processed and sentence tokenized versions for both judgements and summaries. Further, for each sentence in the judgement text we provide a noisy label (75% accurate), which indicates whether or not the sentence is 'summary-worthy'. The strategy for this noisy-labelling is described in our work on legal document summarization [9]. Each judgement and summary sentence is additionally labelled with one of the seven rhetorical roles mentioned in task 1. The *summary-worthy* label as well as the rhetorical roles are assigned automatically and are noisy, which

[2]Team did not submit the working notes
[3]www.nltk.org
[4]www.spacy.io

Team	Run ID	P	R	F	Method Summary
Rustic	rustic_run_1	0.548	0.616	0.557	Legal Bert
MiniTrue	minitrue_run_1	0.485	0.572	0.517	Legal Bert+ simple neural inference network
Arguably	arguably_run_1	0.465	0.591	0.505	Ernie 2.0

Table 1: Results of Task 1: Rhetorical Role Labeling for Legal Judgements.

Team	Run ID	P	R	F	Method Summary
Enigma	enigma_run_1	0.64	0.58	0.59	Legal Bert
NITS	nits_run_2	0.61	0.57	0.58	Legal Bert
NeuralMind	neuralmind_run_1	0.58	0.54	0.54	TextRank

Table 2: Results of Task 2a: Sentence Classification

Team	Run ID	R1	R2	R3	R4	Method Summary
NITS	nits_run_1	0.644	0.363	0.234	0.191	Legal Bert
NeuralMind	neuralmind_run_1	0.629	0.332	0.207	0.153	TextRank
Chandigarh_Concordia	chandigarh_concordia_run_3	0.628	0.338	0.251	0.163	TextRank

Table 3: Results of Task 2b: Summarization

makes this task more interesting as well as challenging. We provide 500, noisily labelled, document-summary pairs for training data. Further 50 documents annotated with rhetorical roles are provided as the test set. Both the train and test sets are part of the dataset produced by our work on legal summarization [9]. For the test dataset we manually annotated the sentences as *summary worthy* or *not summary worthy*.

3.2 Evaluation Methodology

Task 2a is a binary classification task, where the participants are expected to label the sentences as *summary worthy* or *not summary worthy*. For this task we use the standard classification metrics Precision, Recall and F1-Score. We further also report the accuracy, which is percentage of labels predicted correctly. All these metrics are averaged across all documents.

Task 2b is a summarization task where participants can produce an extractive or abstractive summary. For this task we use the standard ROUGE metrics for ranking the submissions. We specifically use ROUGE-1, ROUGE-2 and ROUGE-4 f-scores for this.

3.3 Best Performing Approaches

For both Task 2a and Task 2b we received 11 runs across 5 teams. The best performing systems for Task 2a are listed in Table 2. The runs from team NITS ranked 2nd, 3rd and 4th in the overall rankings, however we only show the best performing run from the team here. While Enigma[5] and NITS[6] both used legal-BERT, Neural mind[5] used the unsupervised TextRank method for generating the sentence ranking.

Most teams directly used the output of Task 2a with an additional constraint on summary length as the summary for Task 2b. The top performing system from NITS used legal-BERT and directly generated an extractive summary. On the other hand Neuralmind and chandigarh-concordia[6] relied on TextRank instead. It was a little surprising to see teams preferring to use an unsupervised algorithm like TextRank in favor of a supervised algorithm that could make use of training data. While in Table 3 the rankings are provided based on ROUGE-1 F-score, we would like to highlight that each Rouge metric is equally important. Using a different ROUGE score for ranking the submissions could easily have created a different rank list.

4 CONCLUSION

The third edition of AILA continued rhetorical segmentation of case documents from last year and offered a new legal summarization task. The results on rhetorical role labelling show a marked improvement over last years' submissions and we see a further scope for improvement. The legal summarization task turned out to be a little challenging for the participants. For the summarization task, we aim to offer a much larger and better annotated datasets in future. This would enable participants make a better use of supervised approaches. We also aim to involve legal professionals for human evaluation of the generated summaries.

REFERENCES

[1] 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* (Kolkata, India).
[2] 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* (Hyderabad, India).
[3] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of Rhetorical Roles of Sentences in Indian Legal

[5]Team did not submit the working notes

[6]Team did not submit the working notes

Judgments. In *Proc. International Conference on Legal Knowledge and Information Systems (JURIX)*.
[4] Sourav Dutta. 2021. Categorizing Roles of Legal Texts via SequenceTagging on Domain-Specific Language Models. In *FIRE 2021 (Working Notes)* (India).
[5] Shaz Furniturewala, Racchit Jain, Vijay Kumari, and Yashvardhan Sharma. 2021. Legal Text Classification and Summarization using Transformers and Joint Text Features. In *FIRE 2021 (Working Notes)* (India).
[6] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of Indian Legal Judgement Documents via Ensembling of Contextual Embedding based MLP Models. In *FIRE 2021 (Working Notes)* (India).
[7] Guneet Singh Kohli, Prabsimran Kaur, and Jatin Bedi. 2021. Automatic Detection of Rhetorical Role Labels using ERNIE2.0 and RoBERTa. In *FIRE 2021 (Working Notes)* (India).
[8] Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. 2017. Overview of the FIRE 2017 IRLeD Track: Information Retrieval from Legal Documents. In *Working notes of Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE) – CEUR workshop proceedings, Volume 2036*. 63–68.
[9] Vedant Parikh, Vidit Mathur, Parth Mehta, Namita Mittal, and Prasenjit Majumder. 2021. LawSum: A weakly supervised approach for Indian Legal Document Summarization. *arXiv preprint arXiv:2110.01188v3* (2021).