TEXAS A&M UNIVERSITY
Engineering

SemEval 2023 Task 6: Rhetorical Roles Prediction (RR): Legal Document Segmentation

Presented By:
Rohan, Hemal, Shubham, Upasana, Abhishek

Problem Statement

TEXAS A&M UNIVERSITY
Engineering

- AI applications in the legal domain such as Judgment summarizing, judgment outcome prediction, precedent search, etc require the segmentation of long and unstructured legal documents into semantically coherent text segments.

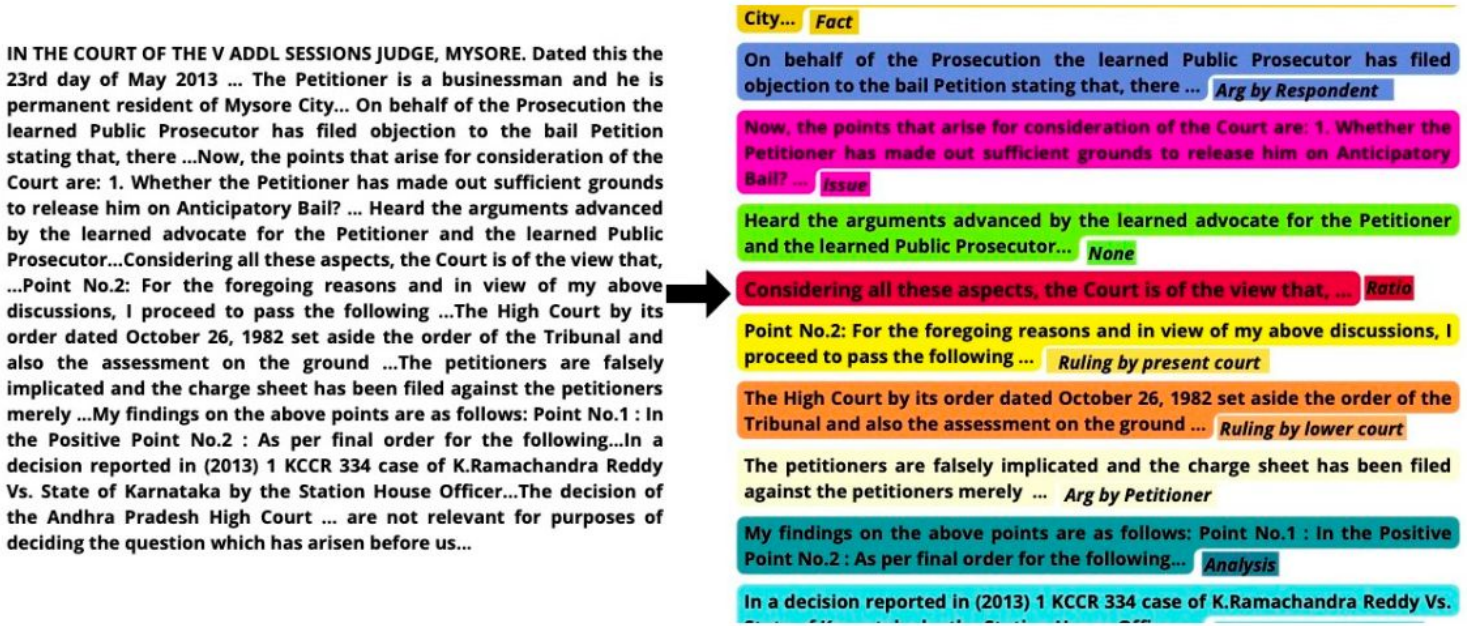
- The problem we aim to solve is to segment a given legal document into topically semantic and coherent units and assign a label to those segments which are referred to as Rhetorical Roles (RR), by predicting the RR label for each sentence.

- The legal document needs to be considered as sequential sentences and the classification of sentences with the prediction of the RR with single label multiple classes for each sentence is needed to be achieved.

Motivation

TEXAS A&M UNIVERSITY
Engineering

- In populous countries (e.g., India), pending legal cases have grown exponentially.
- An automated system can help in many intermediate tasks and expedite the process.



Sub-task B: Legal Named Entities Extraction (L-NER)

Dataset

TEXAS A&M UNIVERSITY
Engineering

- https://github.com/Legal-NLP-EkStep/rhetorical-role-baseline
- Number of documents in Train dataset: 245
- Number of documents in Validation dataset size: 30
- Each document contains sentences which are classified into rhetorical roles
- Total sentences for Training: 28986
- Total sentences for Validation: 2890
- Test dataset will be released in Jan 23

Rhetorical Roles

TEXAS A&M UNIVERSITY
Engineering

1. Preamble: Court name, the details of parties, lawyers and judges' names
2. Facts: Chronology of events
3. Ruling by Lower Court: Judgments given by the lower courts
4. Issues: Such Legal Questions Framed
5. Argument by Petitioner
6. Argument by Respondent,
7. Analysis: Courts discussion on the evidence
8. Statute: discussion on Established laws
9. Precedent Relied: discussions relied
10. Precedent Not Relied: Discussions not relied
11. Ratio of the decision: result of the analysis by the court.
12. Ruling by Present Court: Final decision + conclusion + order of the Court
13. None: Sentence does not fit any of the above

Existing Work

- Past works have centered on the creation of annotated corpora and the task of automatic rhetorical role labeling. Venturi (2012) developed a corpus, TEMIS of 504 sentences annotated both syntactically and semantically. The work of *Wyner et al. (2013)* focuses on the process of annotation and conducting inter-annotator studies.

- One approach was followed using Conditional Random Fields (CRF) with handcrafted features conducted Document segmentation of U.S. court documents to segment the documents into functional and issue-specific parts. Paper *Saravanan et al. (2008)* developed automatic labeling of rhetorical roles, where CRFs were used to label seven rhetorical roles. Bhatia (2014) created Genre Analysis of Legal Texts to create seven rhetorical categories. *Nejadgholi et al. (2017)* developed a method for identification of factual and non factual sentences using fastText.

- Paper Walker et al. (2019) compares different ML approaches and rule-based scripts for rhetorical role identification. The baseline used is a closely related work by paper *Bhattacharya et al. (2019)*, where they label rhetorical roles in Indian Supreme Court documents by using the BiLSTM-CRF model with sent2vec features. Baseline model is a multi-task learning-based model for RR prediction that outperforms the system of paper *Bhattacharya et al. (2019)*.

Experiment Setup

TEXAS A&M UNIVERSITY
Engineering

- The text segmentation and classification problem are modeled as a single-sentence prediction task. Given a legal document D, containing the sentences [s1, s2, ...sn], the task of rhetorical role prediction is to predict the label (or role) yi for each sentence si $\in$ D.

- Multiple approaches have been tried and tested on the given dataset, where different tokenizers and encodings were used to generate embeddings of legal text. These embeddings are then classified as single-label multiclass rhetorical roles using various classifiers. Our final model which surpasses the metrics produced by the baseline model used in the paper follows MultiTask Learning.

Model - Naive Bayes

Naive Bayes classifiers is a family of classifiers with each of them sharing a common principle of all feature pairs being classified being independent from each other. These algorithms work on the base principle of the Bayes' Theorem which can be mathematically stated as follows:

$$P(X|Y) = (P(Y|X) * P(X)) / P(Y)$$

P(X|Y): probability of event X given event Y is true.

P(Y|X): probability of event Y given event X is true.

P(X), P(Y): independent probabilities of events X and Y respectively.

Model - Doc2Vec

TEXAS A&M UNIVERSITY
Engineering

- The main motivation of doc2vec is to represent document into numeric value.
- Doc2Vec model is used to create a vectorized representation of a group of words taken collectively as a single unit, i.e., a document.
- Documents, unlike words, do not possess logical structures. Hence, we add another vector called paragraph ID/ paragraph vector.
- Distributed Memory version of Paragraph Vector (PV-DM) works analogous to the CBOW model of word2vec while Distributed Bag of Words version of Paragraph Vector (PV-DBOW) to the skip-gram model of word2vec.

Model - Wiki BM25 Lexical Model

TEXAS A&M UNIVERSITY
Engineering

- Tokenizer and query encoder of bm25 improved TF-IDF method, with refinements of TF and IDF components is used to generate the embeddings of the legal text.

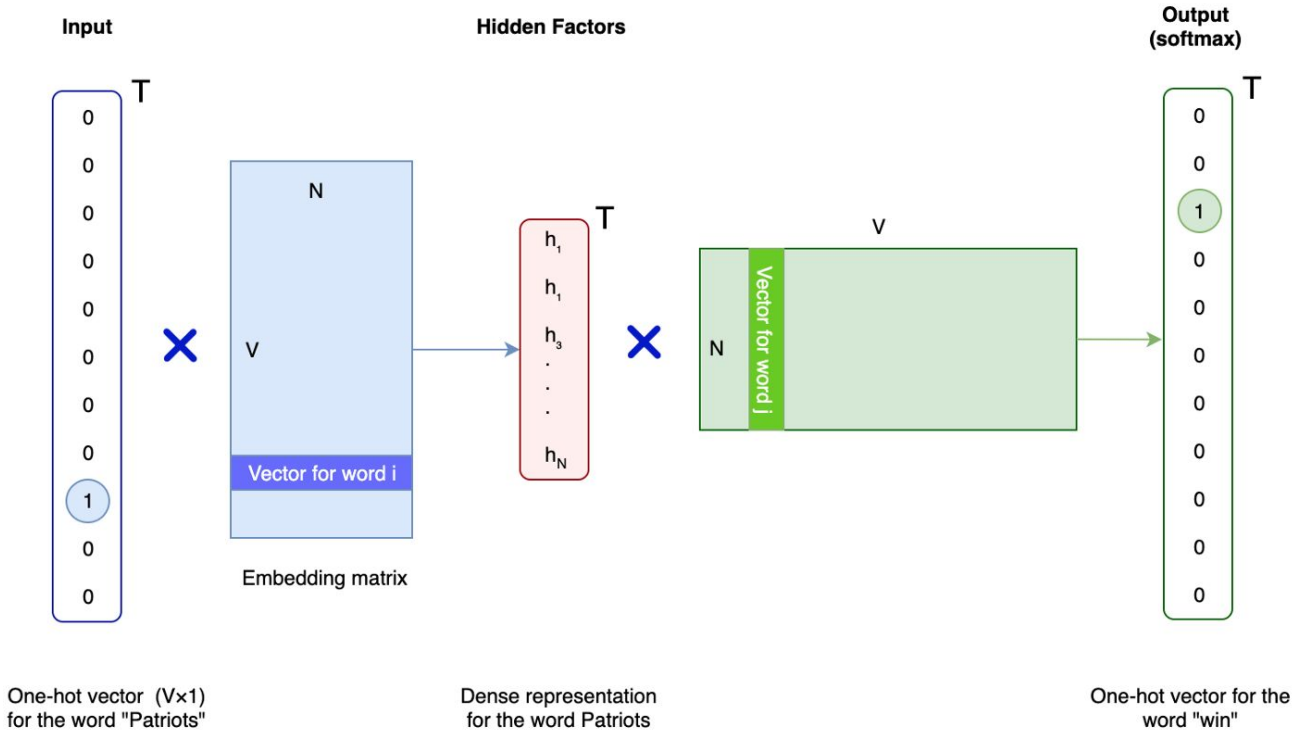
- Classifier Multinomial logistic regression is chosen to perform the classification on word embeddings after training and fitting the model with training data of text embeddings.

- **Analysis**: As multinomial logistic regression is generalized linear model, it cannot fit non-linear data points. The embedding vector data can be seen as a distribution of non-linear points hence, it does not draw a perfect line as a separation between the data points.

Model - Glove Embedding + FCN

TEXAS A&M UNIVERSITY
Engineering

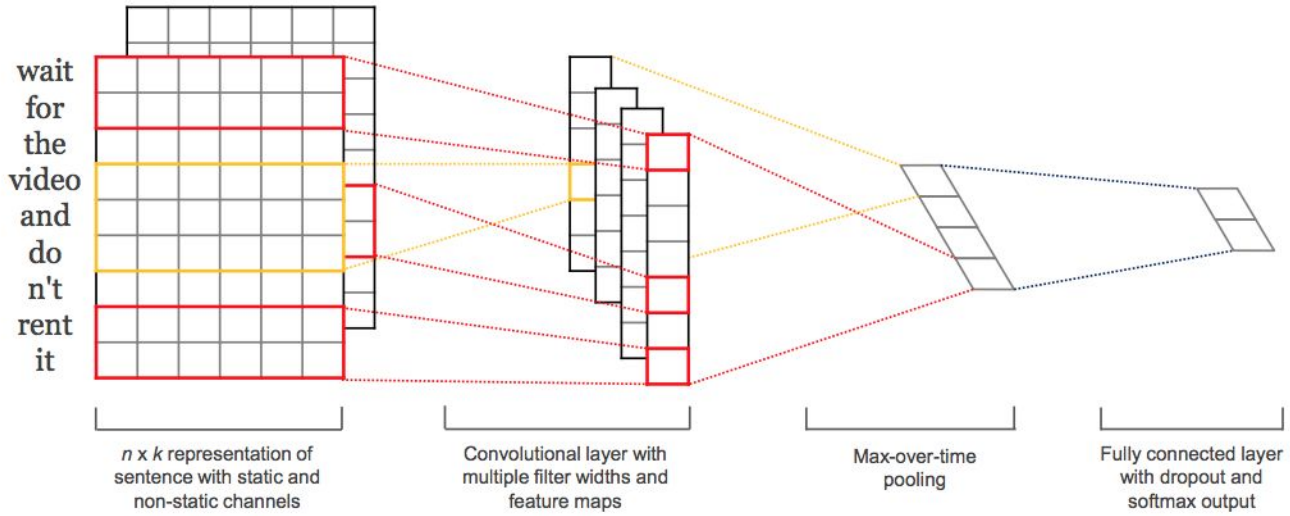
GloVe Embeddings are a type of word embedding that encode the co-occurrence probability ratio between two words as vector differences.



Model - Glove Embedding + 1DCNN

TEXAS A&M UNIVERSITY
Engineering

- convolutional kernel is sliding window, only its job is to look at embeddings for multiple words, rather than small areas of pixels in an image
- similar to representing an n-gram



T5 Model

T5, or Text-to-Text Transfer Transformer, is a Transformer based architecture that uses a text-to-text approach. Every task – including translation, question answering, and classification – is cast as feeding the model text as input and training it to generate some target text. This allows for the use of the same model, loss function, hyperparameters, etc. across our diverse set of tasks. The changes compared to BERT include:

- adding a *causal* decoder to the bidirectional architecture.
- replacing the fill-in-the-blank cloze task with a mix of alternative pre-training tasks.

For our application we have used T5 encoder of several sizes including T5-base, T5-large, T5-xl, and T5-xxl.

T5 Encoder + (FAISS)

TEXAS A&M UNIVERSITY
Engineering

FAISS (Facebook AI Similarity Search) is a library that allows developers to quickly search for embeddings of multimedia documents that are similar to each other. It solves limitations of traditional query search engines that are optimized for hash-based searches, and provides more scalable similarity search functions.

T5 Encoder + (PLDA)

TEXAS A&M UNIVERSITY
Engineering

1. PLDA is a generative model where we assume that the data samples X of a class are generated from a Gaussian distribution. The mean of Gaussian represents the class variable y is generated from another Gaussian distribution called as prior.
1. For the task of recognition, we can compare two examples of unseen classes using PLDA scores by comparing likelihood of examples from same class vs likelihood of examples from different.
2. We can perform clustering of examples into classes using PLDA scores between all pairs of examples from entire set of examples.

T-5 encoder with BiLSTM

TEXAS A&M UNIVERSITY
Engineering

- Used T-5-base model to get embeddings for each sentence and used a bidirectional LSTM model followed by a linear layer to get labels.

```
LSTM(768, 64, num_layers=3, batch_first=True, bidirectional=True)
Linear(in_features=128, out_features=13, bias=True)
```

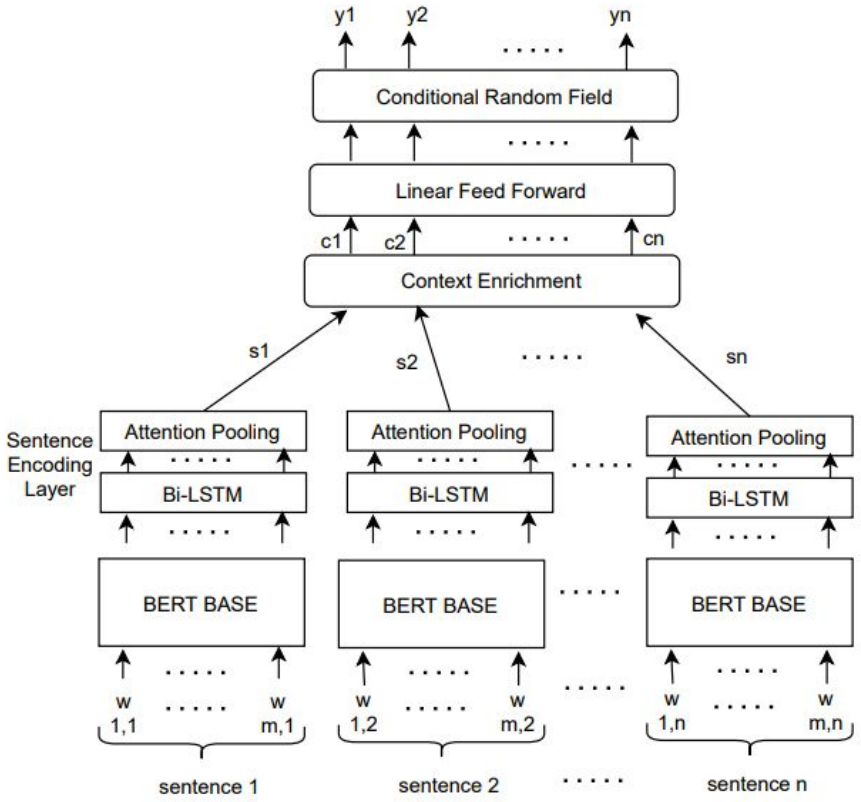
Observations made in Baseline

TEXAS A&M UNIVERSITY
Engineering

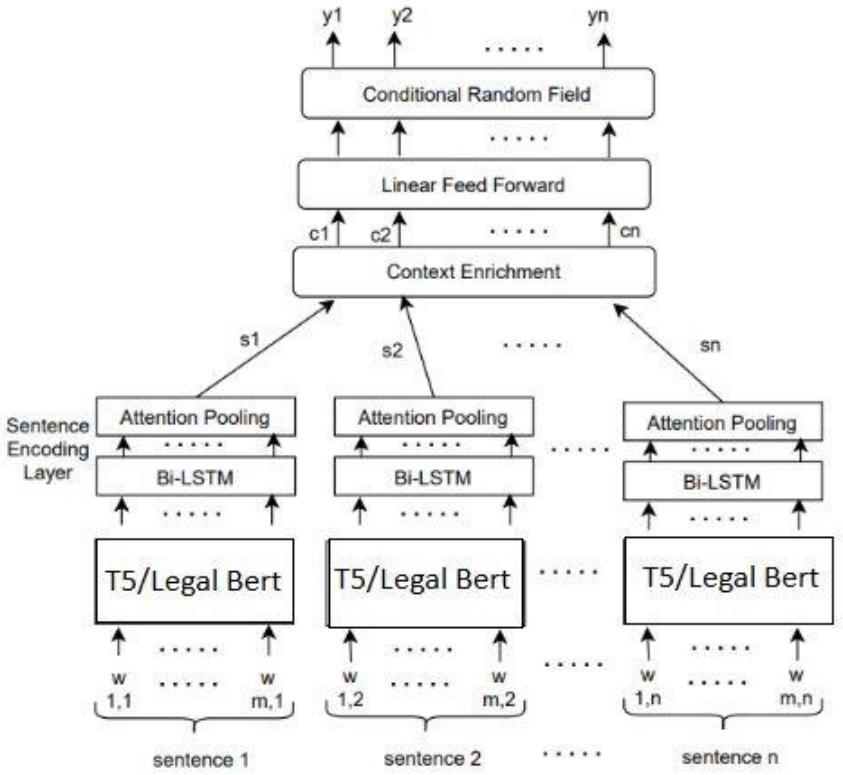
- Rhetorical role labels do not change abruptly across sentences in a document, and the text tends to maintain topical coherence. Given the label y for a sentence s_i in the document, we hypothesize that the chances of shift (change) in the label for the next sentence s_i+1 are low.
- LegalBERT fine-tuned on legal documents performs better than Bert-base-uncased for this task as it is trained on domain specific documents.

Baseline and Our Approach

TEXAS A&M UNIVERSITY
Engineering



Baseline



Our Approach

Data Preprocessing(Baseline)

TEXAS A&M UNIVERSITY
Engineering

The key idea behind the baseline preprocessing approach was to ensure that the sentences of a single document stays in a single batch. This was based on the understanding that the next role is 88% of the times same as that of the previous role.

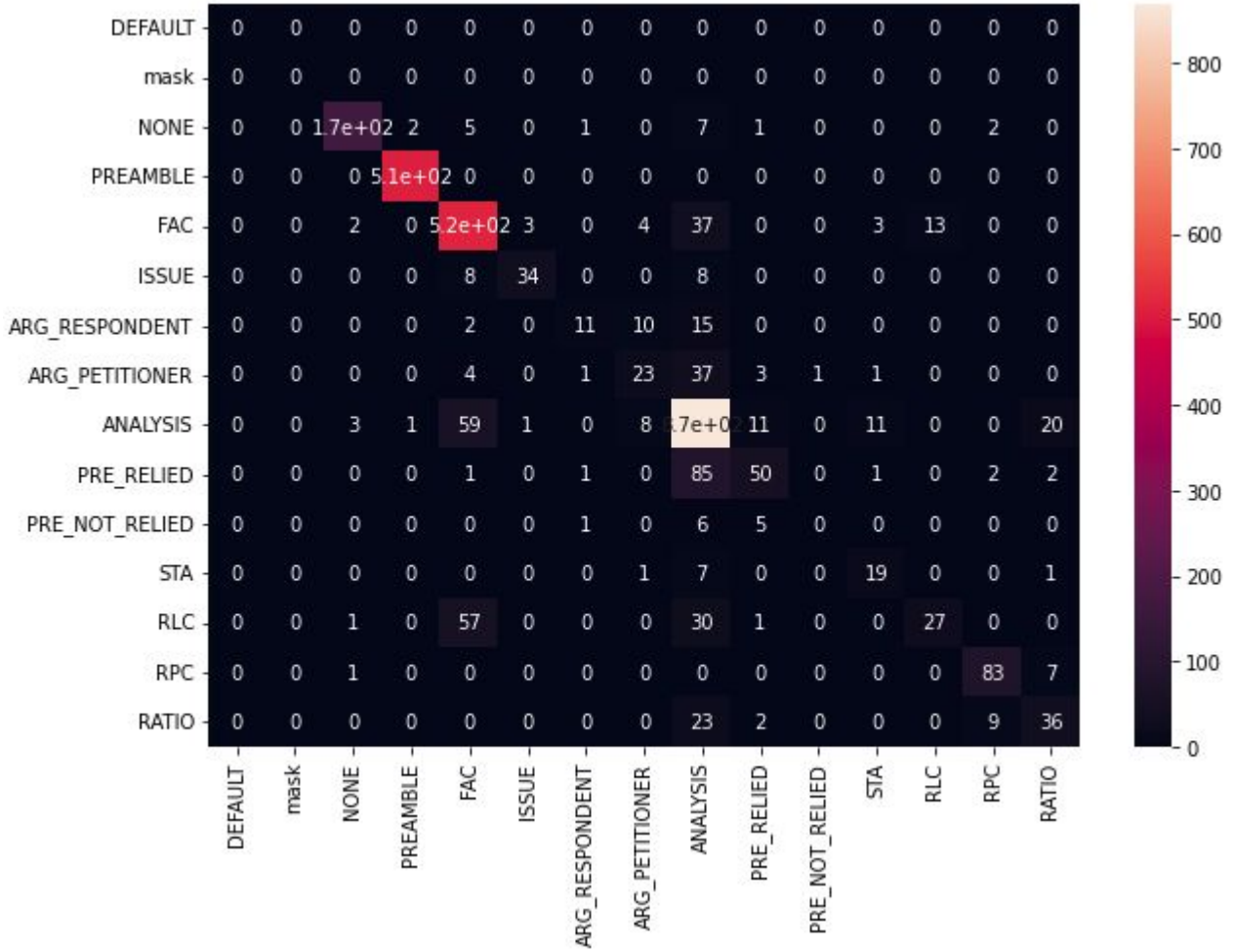
Results

TEXAS A&M UNIVERSITY
Engineering

Model	Macro F1 on Validation	Precision on Validation
Naive Bayes	0.3202008672595919	0.5217426043956311
Doc2Vec	0.35678367302630026	0.46079523508796877
Wiki BM25 Lexical Model + LR	0.382966	0.51591
Glove Embedding + FCN	0.2694453017795601	0.3142378360218918
Glove Embedding + 1DCNN	0.3136997664692862	0.35628853795417964
T5 Encoder + (FAISS)	0.3524619251870554	0.4806228373702422
T5 Encoder + (PLDA)	0.4319042539617168	0.5
T5-base + LSTM	0.3923875427246094	0.4703375424246297
BaseLine	0.5962514492279818	0.8082667592914207
Our Approach (Best) - Baseline modified with T5-large tokenizer and encoder	**0.6161397982433457**	**0.8166029871483154**

Confusion Matrix for Best Model

TEXAS A&M UNIVERSITY
Engineering



Different Configurations

TEXAS A&M UNIVERSITY
Engineering

Model	Macro F1 on Validation	Precision on Validation
Sentence Transformer (paraphrase-xlm-r-multilingual-v1)+ PLDA	0.39319976447977234	0.46401384083044983
Sentence Transformer (paraphrase-xlm-r-multilingual-v1)+ FAISS	0.3329157145830798	0.4598615916955017
Baseline modified with T5-base encoder + BiLSTM with 1024 hidden layer	0.6103711640920119	0.8127822160472387
T5-base encoder + BiLSTM with three hidden layer	0.5796309101679527	0.7964571031608197
Baseline modified with LegalBert	0.587383359920286	0.8089614449461618
Baseline modified with LegalBert Large 1.7m	0.5777515872781539	0.7919416464050018
Our Approach (Best) - Baseline modified with T5-large tokenizer and encoder + BiLSTM	**0.6161397982433457**	**0.8166029871483154**

Analysis

TEXAS A&M UNIVERSITY
Engineering

- Traditional methods like Naive Bayes and embedding based approaches like Glove, Doc2Vec, Wiki-bm25 performed decently.
- Next trials with 1D-Conv and LSTM captured long term dependencies leading to better performance
- By focusing on improving the baseline, T5-large encoder led to best performance.
- Legal-BERT/Legal-BERT-1.7m did not perform better than T5-encoder. This might be attributed to the fact that LEGAL-BERT (trained on EU legal documents, which also has European competition law) is not trained on Indian IT law documents. Using T5-encoder for embeddings with BiLSTM-CRF provides better results.

Summary

We considered the task of classifying sentences in Legal Documents based on their respective rhetorical roles. We tried all kinds of approaches ranging from Naive Bayes, Information Retrieval, LSTMs Large Language Model Embedding. Baseline modified with large language model embedding (T5) outperformed the baseline results.

Limitations

TEXAS A&M UNIVERSITY
Engineering

From each approach:

- TF-IDF is based on the bag-of-words model, therefore it does not capture position in text, semantics, co-occurrences in different documents. It does not capture word semantics in the text hence only useful as a lexical level feature.
- GLOVE embedding does not contain contextual information which could hurt performance.
- 1D Conv and FCN does not capture long range dependencies.
- In our best approach the labels with per-label-precision and per-label-F1 score are Argument by Petitioner and Argument by Respondents are low because they are very similar to each other.
- Also, the model was not able to identify Precedent Not Relied class and classified them as Precedent Relied and Analysis. This can be attributed to incorrect identification of boundary because they are almost similar to one another.

TEXAS A&M UNIVERSITY
Engineering

Q & A

TEXAS A&M UNIVERSITY
Engineering

Thank You

TEXAS A&M UNIVERSITY
Engineering

Appendix

Model Parameters

TEXAS A&M UNIVERSITY
Engineering

"model": "t5-large",
"bert_trainable": false,
"dropout": 0.5,
"word_lstm_hs": 758,
"att_pooling_dim_ctx": 200,
"att_pooling_num_ctx": 15,
"lr": 3e-05, "batch_size": 32,
"max_seq_length": 128,
"max_epochs": 20,
"early_stopping": 5

"weighted-f1": 0.800799221699967 6,
"weighted-precision": 0.806932444734725,

Results Analysis

TEXAS A&M UNIVERSITY
Engineering

"labels": ["DEFAULT", "mask", "NONE", "PREAMBLE", "FAC", "ISSUE", "ARG_RESPONDENT", "ARG_PETITIONER", "ANALYSIS", "PRE_RELIED", "PRE_NOT_RELIED", "STA", "RLC", "RPC", "RATIO"],

"per-label-f1": [0.0, 0.0, 0.932249322493224, 0.997055937193326 7, 0.839546191247974, 0.772727272727272 7, 0.415094339622641 6, 0.396551724137931 05, 0.825035561877667 1, 0.465116279069767 3, 0.0, 0.603174603174603, 0.346153846153846 15, 0.887700534759358 3, 0.529411764705882 2],

"per-label-precision": [0.0, 0.0, 0.960893854748603 3, 0.994129158512720 1, 0.792048929663608 5, 0.894736842105263 2, 0.733333333333333 3, 0.5, 0.773333333333333 3, 0.684931506849315, 0.0, 0.542857142857142 8, 0.675, 0.864583333333334, 0.545454545454545 4],

Confusion Matrix for Best Model

TEXAS A&M UNIVERSITY
Engineering