TeamShakespeare at SemEval-2023 Task 6: Understand Legal Documents with Contextualized Large Language Models

Xin Jin[*]
Northwest Polytechnical Unviersity
xin.jinsw@gmail.com

Yuchen Wang[*]
Ohio State University
wang.9298@osu.edu

Abstract

The growth of pending legal cases in populous countries, such as India, has become a major issue. Developing effective techniques to process and understand legal documents is extremely useful in resolving this problem. In this paper, we present our systems for SemEval-2023 Task 6: understanding legal texts (Modi et al., 2023). Specifically, we first develop the Legal-BERT-HSLN model that considers the comprehensive context information in both intra- and inter-sentence levels to predict rhetorical roles (subtask A) and then train a Legal-LUKE model, which is legal-contextualized and entity-aware, to recognize legal entities (subtask B). Our evaluations demonstrate that our designed models are more accurate than baselines, e.g., with an up to 15.0% better F1 score in subtask B. We achieved notable performance in the task leaderboard, e.g., 0.834 micro F1 score, and ranked No.5 out of 27 teams in subtask A.

1 Introduction

The growing amount of legal cases and documents requires more and more human efforts to process them. (Kalamkar et al., 2022b; Malik et al., 2021) In some countries, such as India, legal cases have accumulated in an incredible number. For example, India has more than 47 million cases pending in the courts (Kalamkar et al., 2022b). This has created a need for automated methods to help judges efficiently understand and process relevant and reliable information. In addition, these methods can also help students, legal scholars, and court officials who deal with legal documents daily. One way to assist them is to automatically understand and highlight the key information and context of long legal documents.

However, understanding legal documents by machines is not an easy task. First, legal documents are often full of technical terminologies, which

can span different legal divisions (Kalamkar et al., 2022b). Furthermore, legal documents can be specific to special cases, such as health (Young, 2009), IT technology (Lu et al., 2017b), and cyber security (Shackelford et al., 2016; Jin et al., 2018; Lu et al., 2017a), which will involve domain-specific words. In addition, legal documents can be extremely long (Kalamkar et al., 2022b), which makes dependency-based techniques and models, such as RNN models, fail to extract the context information due to gradient vanishes. Finally, typos and unstructured documents introduce noises (Kalamkar et al., 2022b), which makes automated natural language processing challenging.

Despite these challenges, predicting rhetorical roles and recognizing named entities in legal documents are very useful for automating the processing and understanding of legal documents. Rhetorical role prediction segments texts and structures noisy legal documents into topically and semantically coherent units (Ghosh and Wyner, 2019). Named entity recognition helps identify key legal entities in long documents (Nadeau and Sekine, 2007), which can not only help judges process cases in a more efficient way, but also benefit the next automation steps. These two tasks can serve as key steps in these methods (Modi et al., 2023).

In this paper, we propose to solve the rhetorical role prediction and named entity recognition problems in the legal document domain with contextualized large-language models. We first systematically build models with well-known design choices based on popular pre-trained models (e.g., BERT and XLM-roBERTa) (Qiu et al., 2020), then systematically evaluate the performance of different models and identify the key limitations, and eventually propose our legal contextualized models as our final solutions. For rhetorical role prediction, we model this task as the sequential sentence classification problem and build the Legal-BERT-HSLN model, which considers the comprehensive

---

[*]These authors contributed equally to this work.

context semantics in both intra- and inter-sentence levels. For named entity recognition, we propose to build a legal-LUKE model that is both sensitive to context and entity-aware. Our evaluation results show that our proposed models are more accurate than baselines, e.g., Legal-LUKE is 15.0% better than our baseline BERT-CRF in F1 score (more details in §4.3). Furthermore, we also achieved the top performance on the rhetorical role prediction task leaderboard, i.e., ranked No.5 out of 27 teams and achieved the 0.8343 micro F1 score (see §4.2).

We briefly summarize our primary contributions as follows.

- We formalize the rhetorical role prediction task as a sequential sentence classification problem and build the Legal-BERT-HSLN framework to model comprehensive sentence semantics.
- We construct the legal-LUKE model with contextualized legal-document and entity-aware representations.
- Our evaluations demonstrate the better performance of our proposed model compared to baselines and achieved promising results on the task leaderboard.

2 Background

2.1 Sequential Sentence Classification

Sequential sentence classification is a natural language processing (NLP) technique that involves classifying a sequence of sentences into one or more categories or labels (Hassan and Mahmood, 2017; Cohan et al., 2019; Jin and Szolovits, 2018; Brack et al., 2022). The objective of this technique is to analyze and classify the content of a given text document based on the semantics and context of the sentences (Jin et al., 2022a). It is often used in classification tasks (Xu et al., 2018; Shan et al., 2021, 2022; Baskaran et al., 2022), such as sentiment analysis, spam detection, and topic classification, where the classification of a single sentence can depend on the preceding or succeeding sentences (Hassan and Mahmood, 2017). This technique can be implemented using various algorithms, such as recurrent neural networks (RNNs) or long-short-term memory (LSTM) networks, which are capable of processing sequential data (Lipton et al., 2015).

In sequential sentence classification, the input is a sequence of sentences and the output is one or more labels that describe the content of the document. The classification can be performed at the sentence level or at the document level, depending on the specific use case (Cohan et al., 2019). Sequential sentence classification is an important technique in NLP because it enables machines to understand and analyze the meaning and context of human language, which is crucial for many applications such as automated text summarization and question-answering systems (Qiu et al., 2020).

2.2 Legal Named Entity Recognition

The objective of named entity recognition in the legal domain is to detect and label all instances of specific legally relevant named entities within unstructured legal reports (Nadeau and Sekine, 2007; Kalamkar et al., 2022a; Li et al., 2020). Using this named entity information, one can analyze, aggregate, and mine data to uncover insightful patterns. Furthermore, the ultimate goal of legal document analysis is to automate the process of information retrieval or mapping a legal document to one or more nodes of a hierarchical taxonomy or legal cases, in which legal NER plays a significant role.

Given that legal reports contain a large number of complex medical terms and terminologies, such as statute and precedent, identifying expressions referring to anatomies, findings, and anatomical locations is a crucial aspect in organizing information and knowledge within the legal domain (Kalamkar et al., 2022a). Therefore, automating this identification process has been recognized as a key target for automation. Moreover, automatic named entity recognition (NER) helps exhaustively extract semantic information and misspelling checking (Kalamkar et al., 2022a; Modi et al., 2023).

3 System Overview

3.1 Rhetorical Role Classification

Identifying rhetorical roles (RR) in legal documents is a challenging task, which requires machine learning models to accurately classify sentences into predefined RR categories. One of the primary challenges is the variability and complexity of natural language, including variations in sentence structure, word choice, and context. This requires training machine learning models on diverse and large datasets that capture the range of languages used in the real world. Another significant challenge is capturing the long-term dependencies between sentences in a sequence, which can
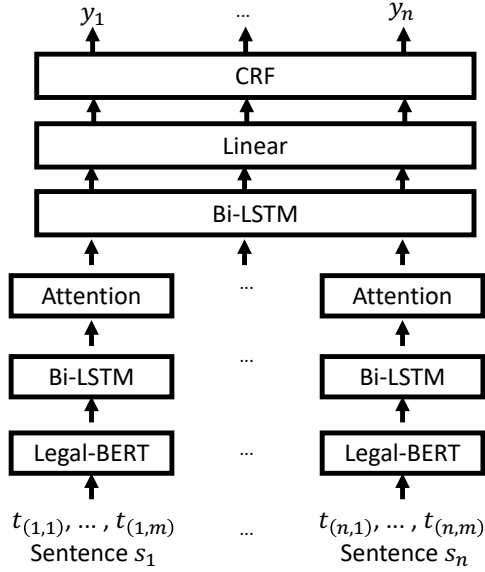
Figure 1: Model Architecture of Legal-BERT-HSLN

be difficult to achieve but is essential to determine the overall meaning and context of the text.

To solve this problem, we first observe that this task is basically a sequential sentence classification problem. Meanwhile, existing work proposes to solve this problem using encoder-decoder models, where the encoder embeds the sentence-level semantics and considers the context dependency information, and the decoder classifies individual sentences by the contextualized surrounding sentence information. For example, Jin et. al (Jin and Szolovits, 2018) classify medical abstract sentences with a hierarchical sequential labeling network, which is composed of a word encoder, a sentence encoder, and a document encoder. Specifically, the word-level encoder is a bidirectional LSTM that encodes each word in a sentence into a vector12. The sentence-level encoder is another bidirectional LSTM that encodes each sentence vector in a hidden state. The document-level encoder is either an LSTM or a CRF layer that models the dependencies between sentences and outputs a label for each sentence.

We follow the paradigm and propose the Legal-BERT-HSLN model for legal rhetorical role classification. In our design, we applied the HSLN structure from Brack et al. (Brack et al., 2022) and changed the backbone model to Legal-BERT (Chalkidis et al., 2020). The model architecture of Legal-BERT-HSLN is shown in figure Figure 1. The model first takes as input the sequences of legal word tokens ($\{t_{i,1}, t_{i,2}, ..., t_{i,m}\}$) of sentence s_i and Legal-BERT generates the correspond-

ing token embeddings. Next, the token embeddings are further enriched with local context information within the sentence s_i by Bi-LSTM and the attention pooling layer as the augmented token embeddings ($\{e_{i,1}, e_{i,2}, ..., e_{i,m}\}$), which is aggregated to generate the embedding of the sentence e_{s_i}. Since one of the most important features in this task is the inter-sentence dependency, we further enrich the sentence embedding with contextual information from surrounding sentences. The output layer transforms contextualized sentence embeddings into rhetorical role labels via a linear transformation and CRF. We also introduce dropout layers after each layer for regularization.

3.2 Legal Named Entity Recognition

In SemEval task 6 (Modi et al., 2023), the second subtask is the recognition of legal entities named entities. Specifically, the legal documents provide nonexhaustive metadata with 14 legal entities, including petitioner, respondent, court, statute, provision, precedents, etc. Identifying these legal entities can be both error prone and labor intensive (Mohit, 2014).

Although the task of legal NER may seem straightforward, it is in fact a challenging undertaking due to several reasons. First, the NER task itself is an unsolved problem (Li et al., 2020). Moreover, legal documents can be very noisy (Kalamkar et al., 2022b,a). Legal texts contain morphological forms (e.g., synonyms, abbreviations, and even misspellings), which means that different legal cases can use different words and phrases to express the same meaning. To process the natural language texts, existing approaches use vocabulary-based embeddings (Wang et al., 2020). However, legal documents involve many out-of-vocabulary words (Jin et al., 2022b), which further makes the task complex.

Meanwhile, we have several insights to solve the problems. First, natural language preprocessing, e.g., tokenization, POS tagging, and sentence parsing, can mitigate the noise of legal documents. Moreover, the identified POS tags and sentence structure can help determine the legal entities by the nature of human languages. Second, compared to static entity representations that assign fixed embeddings to words and entities in the knowledge base, contextualized word representations can generate adaptive semantic embeddings of words and entities, which can be tuned by domain-
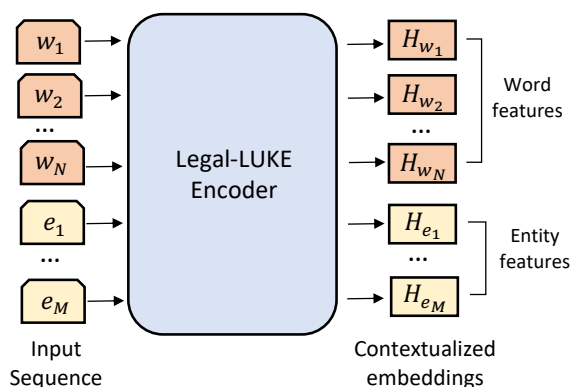
Figure 2: Entity-aware Contextualized Representation

specific context (Ethayarajh, 2019). Finally, we observe that state-of-the-art entity-aware representations (Yamada et al., 2020; Ri et al., 2022) can take advantage of both the advantages of contextualized word embeddings and create entity representations based on the rich entity-centric semantics encoded in the corresponding entity embeddings. Therefore, we combine these insights and propose to identify legal entities with embeddings of legal-sensitive entities.

For legal NER, we introduce a legal entity recognition model: Legal-LUKE, based on the bidirectional transformer encoder of LUKE (Yamada et al., 2020). The LUKE model was pre-trained by predicting both words and entities masked by the [MASK] tokens. As shown in Figure 2, Legal-LUKE takes as input the sequence of preprocessed words $(w_1, w_2, .., w_N)$ and entities $(e_1, e_2, ..., e_M)$. The encoder generates the legal-contextualized representations for both words and entities, i.e., $(H_{w_1}, H_{w_2}, .., H_{w_N})$ and $(H_{e_1}, H_{e_2}, .., H_{e_M})$. To compute the word and entity embeddings, three embeddings are added together, which include token embeddings, type embeddings, and position embeddings. Position embeddings are used to associate entity tokens with their corresponding word tokens, where the position of an entity token is determined by the positions of its corresponding word tokens. The entity position embeddings are then added up over these positions.

4 Evaluations

We conduct extensive experiments using the rhetorical role classification and legal named entity recognition tasks. We have implemented the Legal-BEST-HSLN model and the Legal-LUKE model based on Pytorch (Paszke et al., 2019).

4.1 Experimental Setup

The input word sequence is created by inserting the tokens of [CLS] and [SEP] into the original word sequence as the first and last tokens, respectively, unless stated otherwise. For the input entity sequence of Legal-LUKE, the legal entities from the training set are used to fine-tune the model. Although we cannot access the test set before the task organizers open the evaluation system, we train and tune our models and baselines based on the training and validation set.

4.1.1 Baselines for Rhetorical Role Classification

For subtask A, we set up our experiments by selecting BERT-base and 3 BERT variants with minor modifications as our baselines:

- **BERT-Base**. For this model, we directly use the [CLS] token embedding as the sentence embedding as the encoder of the classifier.
- **BERT-Mean**. Instead of using the hidden state of the token [CLS] for the classification output from BERT, we tried to use the mean value of 128 lengths of sequences.
- **BERT-Regularization**. In this method, we made a preprocessing work to the training set by regularizing symbols. These procedures include: (a) lowercase, (b) remove @name, (c) isolate and remove punctuation except '?', (d) remove special characters such as columns and semi-columns, (e) remove stop words except 'not' and 'can', and (f) remove trailing whitespace.
- **BERT-Augmentation**. Each sentence training set is randomly swapped and the entire training size is doubled.

4.1.2 Baselines for Legal Named Entity Recognition

For subtask B, we select the following 3 models as our baselines:

- **BERT-CRF**. For this model, we use the BERT model as the encoder and transform the encodings into NER labels with the CRF prediction head.
- **BERT-Span**. BERT-Span (Joshi et al., 2020) is the variation of BERT that uses BERT to train span boundary representations, which are more suitable for entity boundary detection.

Model	Micro F1 Score	Best Epoch
BERT-Base	0.631	5
BERT-Mean	0.641	4
BERT-Regularization	0.597	4
BERT-Augmentation	0.645	4
Legal-BERT-HSLN	**0.828**	**16**

Table 1: Rhetorical Role Classification Performance

- **XLM-roBERTa-CRF**. XLM-roBERTa-CRF is a combination of XLM-RoBERTa (Conneau et al., 2019) and CRF. XLM-RoBERTa is a multilingual version of RoBERTa, a transformer model pre-trained on a large corpus in a self-supervised fashion. We used XLM-RoBERTa-large as the encoder model.
- **mLUKE**. mLUKE (Ri et al., 2021) is a multilingual extension of LUKE, a pre-trained language model that incorporates entity information from Wikipedia.

4.1.3 Test Environment

The evaluations are performed on a desktop server, with the Intel Xeon E5-1650 CPU, Ubuntu 18.04 OS, 64 GB memory, 4 TB storage, and 4 NVIDIA GeForce GTX 1080 Ti graphics cards. While training the Legal-LUKE model, we encountered out-of-memory issues, and we switched to another server, which has the Intel Xeon W-2245 CPU, Ubuntu 20.04 OS, 128 GB memory, 2 TB storage, and an NVIDIA RTX A5000 graphics card.

4.2 Results of Rhetorical Role Classification

The goal of rhetorical role classification is to provide a categorization label for each sentence at the unstructured document level. The number of classes is 13. The training set has 247 documents with a total sentence number of 30k, validation set has 30 documents with a total sentence number of 3k. In this section, all evaluations are reported based on the validation set.

Table 1 shows the summary of micro F1 scores for Legal-BERT-HSLN and the baselines on the validation set, where Legal-BERT-HSLN significantly outperforms all baselines and it takes more epochs to converge, where we believe Legal-BERT-HSLN has the better capacity to understand legal documents. It is worth mentioning that the regularization process decreased the F1 score and suggests that the rhetorical role is sensitive to detailed stop words and external sentence markers.

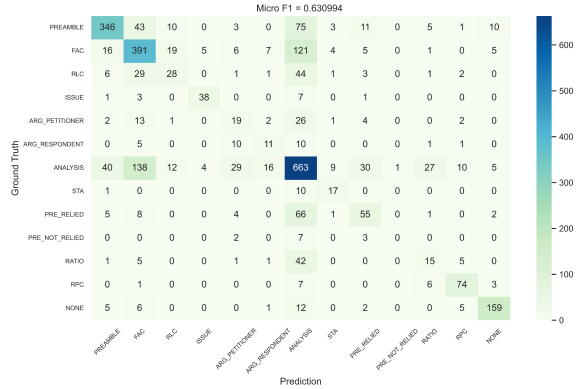While our initial design choice was not Legal-

Figure 3: Confusion Matrix of BERT Baseline Predictions

BERT-HSLN, we originally trained the baseline models. Specifically, our first design choice was to treat all sentences as individual elements where they have no correlation to each other. We set the baseline model with the BERT network backbone with a simple multilayer perceptron. The resulting confusion matrix is shown in Figure 3 with the micro F1 0.631. Subsequently, we conducted 3 BERT variants with minor modifications and obtained the performance shown in Table 1. From the results, we observed the significant performance gap between all the BERT baselines and the state-of-the-art solution (Kalamkar et al., 2022b).

To fully understand the problem, we put an effort into examining the distribution of the legal dataset (Modi et al., 2023) in term of sentence length and classes. In Figure 4 (A), the sentence lengths overall satisfied Poisson's distribution but there is a significant noise at <20 levels. These short sentences may influence the accuracy of classification due to a lack of meaningful information. For example, sentences 17928, 17929, and 17930 are "*[326C-E*", "*] 2.*", and "*There may be circumstances where expenditure, even if incurred for obtaining advantage of enduring benefit would not amount to acquisition of asset*", respectively, whose ground-truth labels are all PREMABLE. This might be due to the discontinuous annotation from turning the page or a keyboard input mistake. Therefore, model prediction from the sentence level is likely not an optimal paradigm to proceed. In Figure 4 (B), we found that there is a significant bias of the labeling in the training set; therefore, we believe that more experiments, such as validating the label bias, can be tested for future research.
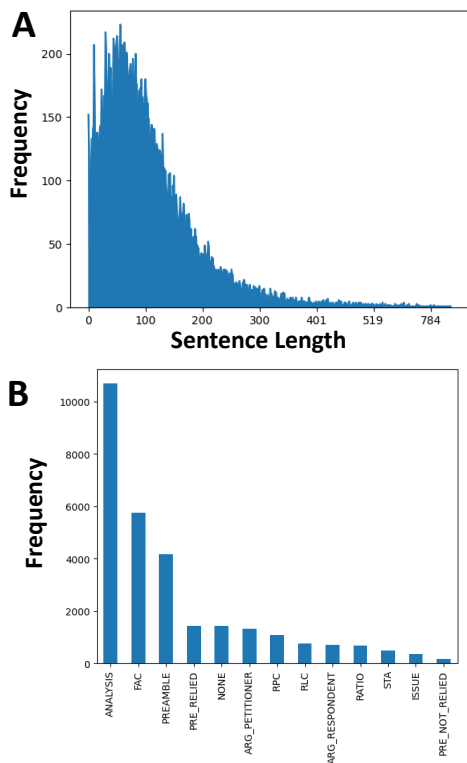
Except for the design choices, we also evalu-

Figure 4: (A) Sentence Length and (B) Class Distributions in Rhetorical Role Dataset
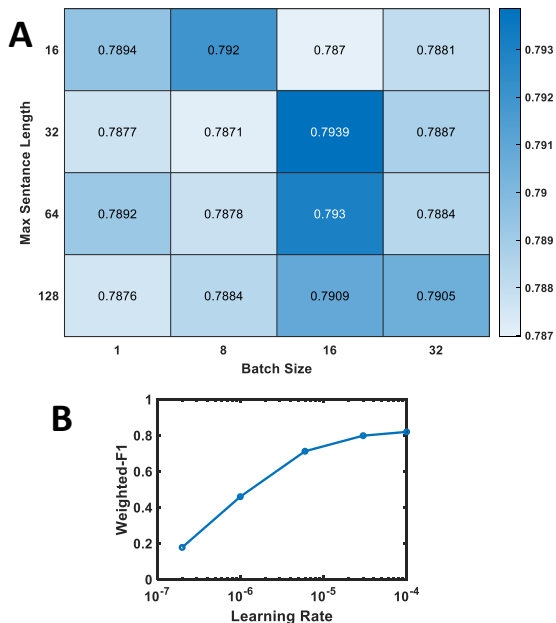

Figure 5: Legal-BERT-HSLN Performance Affected by (A) Maximum Sentence Length and Batch Size and (B) Learning Rate

Figure 6: Confusion Matrix with Legal-BERT-HSLN

ate the effectiveness of our hyper-parameters. Figure 5 shows the micro F1 score is influenced by the batch size, maximum sentence length, and learning rate, where the best combination we have tested so far is batch size 16, maximum sentence length 32, learning rate 1×10^{-4}. All weighted-F1 scores in Figure 5 are calculated from sklearn.metrics.f1_score function with the weighted option. Figure 6 shows the confusion matrix using this combination of hyperparameters, and it is found that the best Micro F1 score we achieve in the validation set is 0.828. Although the major error contribution is no longer in the PREMABLE label, the accuracy of the PRE_NOT_RELIED label is 0 which is the same as reported by the state-of-the-art solution (Kalamkar et al., 2022b). For future experiments, we suggest studying this outlier effect.

Performance on Test Set After the task organizer released the test, we test Legal-BEST-HSLN and submitted the predictions. Surprisingly, we obtained the micro F1 score of 0.8343 and ranked No. 5 out of 27 teams. Note that the test performance is better than all performance on the validation set. Therefore, we believe that there is a shift in the distribution of the training, validation, and test sets.

4.3 Results of Legal Named Entity Recognition

Table 2 presents the overall performance of Legal-LUKE and the baseline models for legal named entity recognition on the validation set. We observe that Legal-LUKE outperforms all baselines with an up to 14.3% better micro F1 score.

Moreover, among the baselines, the BERT-Span model is better than the BERT-CRF model where we used the same encoder model but different pre-

Model	F1 Score	Best Epoch
BERT-CRF	0.694	12
BERT-Span	0.712	14
XLM-roBERTa-CRF	0.773	21
mLUKE	0.787	12
Legal-LUKE	**0.796**	**18**

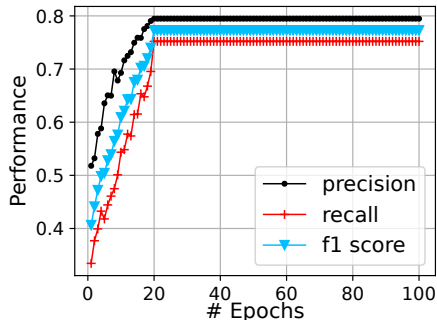Table 2: Overall Performance on Validation Set of Legal Named Entity Recognition

Figure 7: The Best Performance of the Baseline Model XLM-roBERTa-CRF across Training Epochs

diction heads, which demonstrates the benefits of using the span boundary representations for the named entity recognition task. Moreover, the XLM-roBERTa encoder model shows better legal text encoding capacity because the XLM-roBERTa-CRF model and BERT-CRF model share the same decoder module, but XLM-roBERTa-CRF outperforms BERT-CRF with a 11.4% better F1 score. Note that we gave the same parameters for the baseline models. For example, we consider context information by a natural language preprocessor with a maximum length of 100. To balance train efficiency and performance, we set the batch size per GPU as 16 with a learning rate scheduler whose initial rate is 5.0×10^{-6} and the AdamW optimizer. Although we set the maximum number of epochs as 10, we observed that all models achieved the best performance at early epochs. For example, Figure 7 shows the best performance of the XLM-roBERTa-CRF baseline model on the validation set in different epochs, which shows the convergence effect as the training epoch increases.

For our Legal-LUKE model, we also use a learning rate scheduler that has a warm-up ratio of 0.06, the same initial rate 5.0×10^{-6} as baselines, and the AdamW optimizer. While the pre-trained LUKE model is too large for our GPU server, we tune the different batch sizes to trade off between training
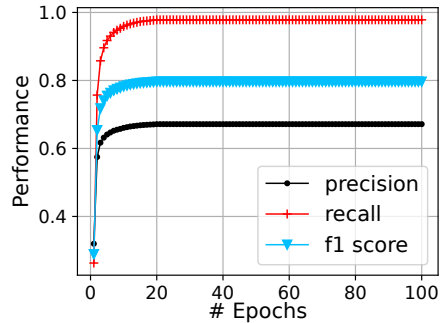

Figure 8: The Best Performance of Legal-LUKE across Training Epochs

speed and performance, and eventually we set the batch size as 8. Similar to the baseline models, we also observe the performance convergence (as shown in Figure 8) though we set the maximum epoch number at 100.

Performance on Test Set After the task organizer released the test, we evaluated Legal-LUKE on the test set and submitted the predictions. We obtained an F1 score of 0.667 and ranked No. 13 out of 17 teams. Our test F1 score is far lower than the validation F1 score. However, according to the dataset (Kalamkar et al., 2022a), the validation set and test set should have similar distributions as they are from the same group of legal cases. Our hypothesis is that the event organizer adopted a different preprocessor from ours which produces different token indices and label shifts. For example, we observed that the legal documents contain many empty lines and we opted to remove such lines in our processing, which can result in the different token indices in our predictions. We plan to further investigate the root cause upon the release of the test set.

5 Conclusion

In this paper, we propose to understand legal documents with context-sensitive language models. For the sub-task of rhetorical role classification, we design the Legal-BERT-HSLN model, which learns the hierarchical context information to solve the sequential sentence classification problem. For legal named entity recognition, we implemented the Legal-LUKE model which is both contextualized and entity-aware. Our evaluation results reveal the outperformance of our models compared to baselines and we are the top-5 teams for the rhetorical role classification task on the leaderboard.

References

Vishal Athreya Baskaran, Jolene Ranek, Siyuan Shan, Natalie Stanley, and Junier B Oliva. 2022. Distribution-based sketching of single-cell samples. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-domain multi-task learning for sequential sentence classification in research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–13.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Saptarshi Ghosh and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, volume 322, page 3. IOS Press.

Abdalraouf Hassan and Ausif Mahmood. 2017. Deep learning for sentence classification. In *2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pages 1–5. IEEE.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. *arXiv preprint arXiv:1808.06161*.

Xin Jin, Wei Lu, Siqi Liu, and Zuqing Zhu. 2018. On multi-layer restoration in optical networks with encryption solution deployment. In *2018 Optical Fiber Communications Conference and Exposition (OFC)*, pages 1–3. IEEE.

Xin Jin, Sunil Manandhar, Kaushal Kafle, Zhiqiang Lin, and Adwait Nadkarni. 2022a. Understanding iot security from a market-scale perspective. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1615–1629.

Xin Jin, Kexin Pei, Jun Yeon Won, and Zhiqiang Lin. 2022b. Symlm: Predicting function names in stripped binaries via context-sensitive execution-aware code embeddings. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1631–1645.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022a. Named entity recognition in indian court judgments. *arXiv preprint arXiv:2211.03442*.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022b. Corpus for automatic structuring of legal documents. *arXiv preprint arXiv:2201.13125*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Zachary C Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Wei Lu, Xin Jin, and Zuqing Zhu. 2017a. Game theoretical flexible service provisioning in ip over elastic optical networks. In *2017 Opto-Electronics and Communications Conference (OECC) and Photonics Global Conference (PGC)*, pages 1–3. IEEE.

Xiaofeng Lu, Ruonan Zhang, Yuliang Zhou, Jiawei Liu, Xin Jin, Qi Guo, and Chang Cao. 2017b. Convolutional modeling and antenna de-embedding for wideband spatial mmwave channel measurement. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2021. Semantic segmentation of legal documents via rhetorical roles. *arXiv preprint arXiv:2112.01836*.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Guha, Sachin Malhan, and Vivek Raghavan. 2023. SemEval-2023 Task 6: LegalEval: Understanding Legal Texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics (ACL).

Behrang Mohit. 2014. Named entity recognition. *Natural language processing of semitic languages*, pages 221–245.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2021. mluke: The power of entity representations in multilingual pretrained language models. *arXiv preprint arXiv:2110.08151*.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Scott J Shackelford, Scott Russell, and Andreas Kuehn. 2016. Unpacking the international law on cybersecurity due diligence: Lessons from the public and private sectors. *Chi. J. Int'l L.*, 17:1.

Siyuan Shan, Vishal Athreya Baskaran, Haidong Yi, Jolene Ranek, Natalie Stanley, and Junier B Oliva. 2022. Transparent single-cell set classification with kernel mean embeddings. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Siyuan Shan, Yang Li, and Junier B Oliva. 2021. Nrtsi: Non-recurrent time series imputation. *arXiv preprint arXiv:2102.03340*.

Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. 2020. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, 11:1611–1630.

Yan Xu, Siyuan Shan, Ziming Qiu, Zhipeng Jia, Zhengyang Shen, Yipei Wang, Mengfei Shi, I Eric, and Chao Chang. 2018. End-to-end subtitle detection and recognition for videos in east asian languages via cnn ensemble. *Signal Processing: Image Communication*, 60:131–143.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Andy Young. 2009. The legal duty of care for nurses and other health professionals. *Journal of Clinical Nursing*, 18(22):3071–3078.