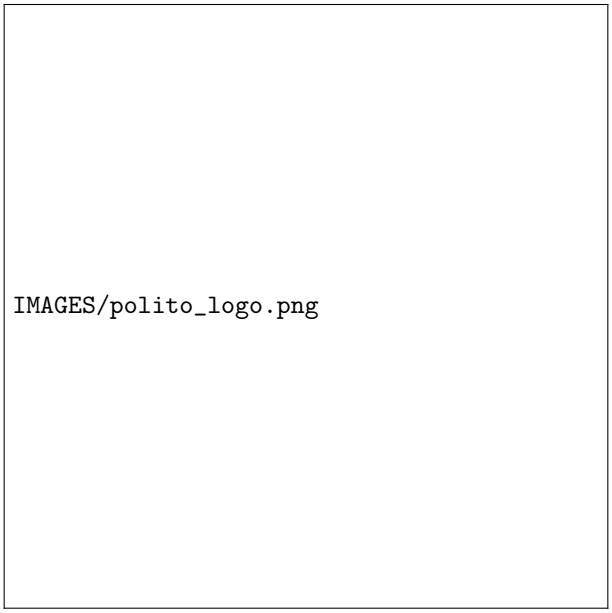

Deep Natural Language Processing

Roxane GOFFINET, Neralb CIKA, Alejandro MESA GOMEZ and Vida AHMADI

January 2024

Named Entity Recognition

Supervised by Lorenzo VAIANI

FOR US

Project tutoring

- 22/12/2023 from 1pm to 4pm ONLINE (Virtual Classroom BBB)
- 12/1/2024 from 2:30pm to 4pm ONLINE (Virtual Classroom BBB)
- 16/1/2024 from 4pm to 7pm Room 29 + ONLINE (VirtualClassroom BBB)

Goal

- Understanding of the assigned paper
- Suggestions/recommendation on data retrieval and preparation, methodology, experiments, validation, method extensions
- No code debugging

Ideas and Guidelines

three specific research lines in Legal AI: Named Entity Recognition, Court judgment Prediction, and Predictions Explanation [9].

to remember :

difference between large and base models : same architecture but higher number of attention layers in huge and also the size of the embedding is bigger.

Multi : means that it manages document with one sentence or multiple sentences.

Extensions: propose new models, propose new training dataset, evaluate the impacts of parameters etc.

Acknowledgements

This report and the research behind it would not have been possible without the support of our supervisor, Lorenzo Vaiani. We would also like to thanks the professor Luca Cagliero along with his team for the quality of the teaching in the Deep Natural Language Processing course. We really enjoyed working on this project and this course in general.

To enhance the clarity and coherence of our report, especially considering that English is not our native language, we leveraged ChatGPT to rewrite certain sections. This approach helped ensure that the content adheres to standard English conventions and effectively communicates our research findings. However, all the research conducted, as well as the results generated and the code developed, are the outcome of our personal efforts.

1 Introduction

This research draws from the insights presented in two key articles: *"Indian Legal Documents Corpus for Court Judgment Prediction and Explanation"* [8] and *"Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction"* [9]. The primary objective is to optimize the preprocessing techniques, refine the dataset, and employ advanced models to achieve superior Named Entity Recognition (NER) results specifically tailored for legal data.

NER remains a well-explored challenge in Natural Language Processing (NLP), supported by various publicly available pre-trained models. However, legal documents pose unique complexities due to their distinctive named entities such as petitioner names, respondent details, court references, statutes, provisions, precedents, and more. These specialized entity types are not effectively identified by conventional Named Entity Recognizers like spaCy. Therefore, there is a critical necessity to develop NER models that are finely attuned to the nuances of legal terminology and structures, especially as it holds significant importance serving as a foundational step for various crucial tasks in the legal domain, including automated summarization, legal recommendation systems, and comprehensive case law analysis. The accuracy and reliability of NER models specifically designed for court judgment texts are pivotal in ensuring the efficiency and effectiveness of downstream applications in the legal realm.

In this research endeavor, our primary objective is threefold. Initially, we aim to replicate the findings and performance metrics of the NER model proposed in the article *"Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction"* [9]. This step will serve as a benchmark for our subsequent improvements and modifications. Subsequently, we will introduce enhancements to the existing model, focusing on refining methodologies, exploring alternative techniques, or leveraging novel approaches to achieve superior NER outcomes for legal entity recognition. Our aim is to investigate and showcase the potential for enhancing the model's performance in accurately identifying and classifying legal entities within judgment texts. Finally, our research culminates in a comprehensive evaluation and analysis of the developed models. We will critically assess their strengths and weaknesses, highlighting the nuanced intricacies of NER models tailored specifically for legal recognition. Moreover, our conclusions will provide invaluable insights into the efficacy of these models, offering guidance and recommendations for future advancements in this domain.

2 Methods

Within the realm of Named Entity Recognition (NER) for legal documents, diverse models have been explored and referenced extensively in the pivotal works [8, 9], as for example BERT, RoBERTa, and LUKE among others. Earlier attempts were based on the use of models such as Hidden Markov Models (HMM) [2] and Conditional Random Field (CRF) [1]; however, their performance, as highlighted by Au, Lampos, and Cox (Dec. 2022)[7], fell short in comparison to the efficacy demonstrated by LUKE and BERT [7]. As a result, our methodological approach focuses on the more promising models BERT, RoBERTa, and LUKE leveraging their potential for Named Entity Recognition in legal texts. A brief description of these models is provided in the appendix.

Given the scarcity of annotated legal datasets and the structure of the model used, most models have undergone pre-training on general English corpora. Yet, the importance of fine-tuning these models on legal data stands out prominently, as elucidated by Au, Lampos, and Cox (Dec. 2022)[7]. Their research emphasizes the necessity of fine-tuning on legal datasets, as the performance of general NER models without this specialized adaptation yields suboptimal results within the legal domain. Thus, our methodology entails this crucial phase of fine-tuning on legal datasets to optimize model performance. The models presented in Benedetto et al.'s article were trained with the Legal-NLP-EKStep dataset that can be found here. This

dataset is composed of two files, preamble and judgement. It has been made specifically for tackling legal-NER Task and to train model with the peculiarity of legal processes and the terminoligies used. The corpus is composed of 9435 judgement sentences and 1560 preambles of Indian court judgment annotated with 14 legal named entities.

The evaluation of these models will be conducted by comparing their F1 scores on the designated test set. This meticulous assessment using F1 scores serves as a robust benchmark for gauging the efficacy and suitability of these models in the context of Named Entity Recognition for legal documents. The F1 score stands as a fundamental metric in evaluating Named Entity Recognition tasks due to its ability to balance precision (the ratio of true positive entities to all entities predicted as positive) and recall (the ratio of true positive entities to all actual positive entities) effectively. F1 score consolidates these two metrics into a single value : their harmonic mean, the formula of the F1 score is given by : $F_1 = \frac{2}{\text{recall}^{-1}+\text{precision}^{-1}} = 2\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$. The F1 score's ability to provide a balanced evaluation, considering both false positives and false negatives, renders it a widely accepted and utilized metric in assessing the effectiveness of NER models. We will distinguish different F1 scores :

- F1 Strict: exact boundary surface string match and entity type;
- F1 Exact: exact boundary match over the surface string, regardless of the type;
- F1 Partial: partial boundary match over the surface string, regardless of the type;
- F1 Type: some overlap between the system tagged entity and the gold annotation is required;

3 Results

3.1 Reproducibility

Our initial aim was to employ the models outlined in the papers and juxtapose our findings with theirs. We utilized the code made available on their GitHub pages for this purpose. However, we encountered uncertainty regarding the methodology employed for dataset preparation and segmentation in their studies. To address this, we independently partitioned our dataset into train, development, and test sets. Consequently, there may be slight discrepancies in the results due to this divergence in dataset handling. Nonetheless, our results closely mirror those reported in the articles, with the notable exception of the LegalRoBERTa-base model. This discrepancy could potentially stem from the model's sensitivity to training data, leading to considerable variability in results when dataset compositions differ.

It is worth noting that training the models used in this study demanded considerable computational resources and time investment. Several of these models necessitated extensive training periods, with some exceeding 24 hours to converge to satisfactory performance levels. This substantial time requirement underscores the complexity and resource-intensive nature of training large-scale language models, especially when fine-tuning them on domain-specific datasets like legal texts. Our results as well as the ones presented in *"PoliToHFI at SemEval-2023 Task 6: Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction"* [9]

Models	Results of the article on the dev set				Our results on the dev set			
	F1 S	F1 P	F1 E	F1 TM	F1 S	F1 P	F1 E	F1 TM
BERT-large (ft on NER)	83.96%	89.64%	85.37%	90.95%	84.38%	89.69%	85.52%	91.11%
RoBERTa-large (ft on NER)	88.38%	92.80%	89.63%	93.56%	89.82%	93.52%	90.82%	93.99%
LegalBERT-base	87.76%	92.15%	88.70%	93.41%	87.68%	92.08%	88.74%	93.33%
LegalRoBERTa-base	86.39%	91.36%	87.66%	92.47%	79.4%	85.32%	80.15%	88.35%
BERT-base (ft on EURLEX)	86.34%	91.66%	87.83%	92.53%	86.33%	91.29%	87.54%	92.03%
BERT-base (ft on ECHR)	86.77%	91.76%	88.13%	92.65%	86.77%	91.57%	87.8%	92.92%
LUKE-base	88.89%	92.73%	89.85%	93.49%	88.9%	92.59%	89.76%	93.32%
LUKE-large	89.88%	93.45%	90.68%	94.20%	90.02%	93.62%	90.96&	94.16%

Table 1: Comparison of the results

3.2 Extentions

3.2.1 Pretraining LUKE model on the EURLEX Dataset

For our first extension we aimed to further pretrain the LUKE model using the EurLex dataset [10], a public corpus containing annotated legal documents. The pretraining process involved leveraging the masked language modeling objective to enhance the model's ability to understand legal text and recognize named entities within it. We chose in particular the EurLex dataset due to its rich annotations and relevance to legal NER tasks. Especially, pretraining on the EURLex dataset will allow the model to capture domain-specific knowledge present in legal documents. This enhances its ability to understand legal text and recognize named entities within it, leading to improved performance on legal NER tasks. Furthermore, LUKE's architecture enables it to understand both textual and non-textual elements in the input. By pretraining on legal text, the model can better contextualize entity mentions within the broader legal context, leading to more accurate entity recognition. And lastly, pretraining the LUKE model on a large corpus like EurLex provides a strong initialization point for further fine-tuning on downstream tasks. That means that we would need smaller amounts of task-specific data.

Due to the unsupervised nature of pretraining, direct model evaluation on a specific task was not conducted in this phase. However, the quality of the pretrained LUKE model could be assessed in the downstream task that is legal NER. The results of the NER task are presented in the Table **??**.

Models	F1 S	F1 P	F1 E	F1 TM
LUKE-base Pretrained on EURLex	88.13%	92.01%	88.85%	93.30%
LUKE-base	88.90%	92.59%	89.76%	93.32%
LUKE-large Pretrained on EURLex	89.84%	93.04%	90.35%	93.87%
LUKE-large	90.02%	93.62%	90.96%	94.16%

Table 2: Comparison of NER score for LUKE model with and without pre-training

As observed, there appears to be no discernible difference between LUKE with and without pretraining on the EurLex dataset. However, due to the extensive runtime required – approximately 4 days for 5 epochs – we were unable to conduct sufficient training. Typically, we anticipate at least 40 epochs for comprehensive training, which may explain the initially underwhelming results. It is plausible that further training iterations could lead to performance enhancements. Additionally, potential inconsistencies or errors in entity annotations within the EurLex dataset might have introduced noise during pretraining, potentially undermining the model's performance.

3.2.2 Data-augmentation

Our second extension idea involved employing data augmentation techniques to expand our dataset for training, with the aim of potentially achieving improved results. This concept was inspired by the article titled *"An Analysis of Simple Data Augmentation for Named Entity Recognition"*[5], where four techniques are introduced for generating new annotated data from existing annotations. From these techniques, we implemented two: Label Wise Token Replacement and Shuffle within Segments.

Label-wise token replacement (LWTR) involves replacing tokens in a sequence while maintaining the original label sequence.This process utilizes a binomial distribution to determine wether a token should be replaced.If replacement is chosen, a label-wise token distribution, derived from the original training set, is utilized to select a replacement token with the same label. As a result, the label sequence remains unchanged.

While Shuffle within segments (SiS) involves dividing the token sequence into segments based on their labels. Each segment represents either a mention or a series of out-of-mention tokens. Then for each segment, a binomial distribution is utilized to determine whether it should undergo shuffling. If shuffling is chosen, the tokens within the segment are reordered while maintaining the original label sequence.

We implemented this technique using the Indian legal dataset mentioned earlier, as the dataset provided in the article was in German. To assess its efficacy, we applied it to 50% of the preamble data. It's worth highlighting that direct comparison with previously reported scores isn't feasible due to the considerable difference in dataset sizes. Nevertheless, we can examine the impact of data augmentation relative to the model trained on the same dataset. In Table **??**, we'll compare the outcomes of data augmentation across BERT (general-purpose), BERT (trained for general NER tasks), and LUKE.

Models	Baseline			LWRT			SiS		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
BERT	39.62%	15.63%	22.42%	41.62%	18.00%	25.13%	40.39%	16.89%	23.83%
BERT NER	42.39 %	16.40%	23.65%	42.50%	18.08%	25.36%	41.53%	16.62%	23.74%
LUKE	44.09%	14.79%	22.15%	44.09%	14.79%	22.15%	42.25%	15.04%	22.19%

Table 3: Comparison of the Data augmentation results

As evident from the results, data augmentation techniques show promise across all three models. While there isn't significant variation in precision, there is noticeable improvement in recall and consequently, the F1 score, particularly with LWRT. This leads us to confidently assert that data augmentation could play a crucial role in enhancing the performance of legal NER, especially in scenarios where annotated data is limited.

4 Conclusions

In conclusion, our experimentation with various techniques aimed at improving Legal-NER yielded mixed results. While fine-tuning LUKE on another dataset did not produce significant improvements, this may be attributed to the insufficient number of epochs run during training. Encouragingly, we obtained satisfactory results with data augmentation, prompting consideration for further exploration, including methods such as synonym replacement or mention replacement. However, due to challenges encountered with applying synonym replacement to languages other than German and the perceived limited relevance of mention replacement for our task, we did not delve deeply into these approaches. Additionally, we introduced FLAIR, a model developed by Humboldt University of Berlin, renowned for its performance across various NLP tasks including NER. Regrettably, we were unable to obtain results with FLAIR due to resource constraints, in-

cluding limited access to GPUs and time constraints. Furthermore, we contemplated leveraging the EDGAR database, as referenced in the article "E-NER — An Annotated NamedEntity Recognition Corpus of Legal Text" [7] published in 2022. We believe that these avenues of exploration will provide valuable insights into the potential of these techniques in enhancing Legal-NER.

References

[1] Andrew McCallum and Wei Li. "Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 188–191. URL: https://aclanthology.org/W03-0430.

[2] Sudha Morwal, Nusrat Jahan, and Deepti Chopra. "Named Entity Recognition using Hidden Markov Model (HMM)". In: *International Journal on Natural Language Computing* 1 (2012), pp. 15–23. DOI: 10.5121/ijnlc.2012.1402.

[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[4] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[5] Xiang Dai and Heike Adel. "An Analysis of Simple Data Augmentation for Named Entity Recognition". In: *International Conference on Computational Linguistics*. 2020. URL: https://api.semanticscholar.org/CorpusID:225041226.

[6] Ikuya Yamada et al. *LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention*. 2020. arXiv: 2010.01057 [cs.CL].

[7] Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. "E-NER — An Annotated Named Entity Recognition Corpus of Legal Text". In: *Proceedings of the Natural Legal Language Processing Workshop 2022*. Ed. by Nikolaos Aletras et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 246–255. DOI: 10.18653/v1/2022.nllp-1.22. URL: https://aclanthology.org/2022.nllp-1.22.

[8] Prathamesh Kalamkar et al. *Named Entity Recognition in Indian court judgments*. 2022. arXiv: 2211.03442 [cs.CL].

[9] Irene Benedetto et al. "PoliToHFI at SemEval-2023 Task 6: Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1401–1411. DOI: 10.18653/v1/2023.semeval-1.194. URL: https://aclanthology.org/2023.semeval-1.194.

[10] *EUR-Lex dataset*. Accessed: 2024-02-15. URL: %5Curl%7Bhttps://huggingface.co/datasets/eurlex%7D.

5 Appendix

5.1 Description of the main models

5.1.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art transformer-based language model developed by Google [3]. This revolutionized the world of NLP by pre-training a deep neural network on massive amounts of text data, thus capturing bidirectional context and producing contextualized word embeddings. BERT uses a transformer architecture, consisting of multiple encoder layers, that enables the learning of intricate patterns and relationships between words in a sentence. It also uses a masked language model (MLM) pre-training objective, where it masks certain words in a sentence and trains the model to predict these masked words based on their context within the sentence. For NER tasks, BERT can be used in a fine-tuning process where the pre-trained BERT model is further trained on labeled NER datasets specific to the legal domain. This fine-tuning involves updating BERT's parameters using the annotated legal data to adapt the model's understanding to recognize named entities such as names of people, organizations, locations, statutes, or legal concepts. During fine-tuning for NER, the input text sequences are tokenized and fed into the BERT model, which assigns a label to each token indicating its entity type (e.g., PERSON, ORGANIZATION, LOCATION). The fine-tuned BERT model learns to predict these entity labels based on the contextual information it has acquired through pre-training and subsequent domain-specific training.

5.1.2 RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) is a variant of the BERT model developed by Yinhan Liu and Myle Ott from Facebook AI [4]. RoBERTA model is designed to enhance the pre-training process and improve language understanding compared to the original BERT architecture. As BERT, it also employs a transformer architecture, utilizing multiple layers of encoders to capture bidirectional context in text and adopts a similar masked language model pre-training objective where it predicts masked words within sentences. However, RoBERTa refines the pre-training methodology by using larger batch sizes, more training data, and longer training epochs. It removes the next sentence prediction objective used in BERT and applies dynamic masking strategies to enhance the model's ability to learn from diverse contexts in text.

For NER tasks, RoBERTa is used similarly to BERT. The model can undergo fine-tuning on labeled legal datasets to adapt its understanding of context and improve its ability to recognize named entities within legal documents.

5.1.3 LUKE

LUKE (Language-Understanding Knowledge-Enhanced) is a model that was proposed in 2020 by [6]. It is tailored for incorporating both linguistic information and knowledge base representations into contextualized embeddings and aims to bridge the gap between language understanding models like BERT and factual knowledge stored in knowledge bases. Unlike BERT and RoBERTa, LUKE employs an entity-aware architecture that encodes information about named entities directly into the model's embeddings during pre-training. This model architecture enables LUKE to generate embeddings that contain not only contextualized information from text but also knowledge graph information associated with named entities (coming from Wikipedia pages for example).

For NER tasks, LUKE can be used similarly to BERT and RoBERTa. The strength of LUKE lies in its ability to embed entities directly into the model's representations, thus potentially enhancing the model's performance in NER tasks by explicitly leveraging factual knowledge about entities.

6 TO REMOVE

6.1 Data

There are not many publicly available annotated datasets for NER task. But general NER is not sufficient to provide great results due to peculiarity of legal processes and the terminoligies used. That is why it is important to develop seperate legal NER for court judgment texts. The data used for the training in the articles were accesible, it was the indian annotated dataset (train and dev) with preamble and judgement that can be found here.

The corpus is composed of 9435 judgement sentences and 1560 preambles of Indian court judgment annotated with 14 legal named entities. Since we didn't have access at the split dataset (train, dev, test), we had to reconstruct the dataset and divide it into a train, dev and test set. In on of our exploration we also used the legal NER dataBase called EDGAR wich is mentionned in the article "*E-NER — An Annotated Named Entity Recognition Corpus of Legal Text*" [7] written in 2022. We also looked into other Legal-Dataset for NER tasks such as a german Named entity recognition dataset that can be found here. This dataset consists of 66,723 sentences with 2,157,048 tokens. The sizes of the seven court-specific datasets varies between 5,858 and 12,791 sentences, and 177,835 to 404,041 tokens. But sadly since the languages of these dataset is german it would take quite a long time to transform the dataset into one that could be usable for our models and it would also require implementing a model for traduction. And while it could have been reaaly interesting to dig deeper in this, we saddly didn't have enough time to explore this option.

There's not a lot of publicly available data sets for doing named entity recognition tasks, especially in the legal field. General NER models often don't work that well on legal texts since they use real specialized language and complicated ideas and so it's important to build NER models just for court cases and judgments that understand all the special legal terms they use. The data sets talked about here come from an Indian data set of court judgments with labels for different types of legal entities. It has about 9000 sentences from Indian court cases annotated with 15 different types of legal named entities. Since they didnt already split it into train, dev and test sets, we had to reconstruct those by dividing it ourselves into separate training development and test data.

During our exploration, we used the EDGAR legal named entity recognition (NER) database from the 2022 article E-NER - An Annotated Named Entity Recognition Corpus of Legal Text. We also looked at other legal datasets for NER, like a German dataset available online. This German dataset has over 66,000 sentences and 2 million tokens while other court datasets range from 5,000-12,000 sentences. However, these are in German so using them would need translation models which takes time. While these could give good insights, we couldn't research them further due to time limits.