Legal Named Entity Recognition

1st Farhad Yousefi Razin
dept. Informatica
Politecnico di Torino
s310027@studenti.polito.it

2nd Shayan Bagherpour
dept. Informatica
Politecnico di Torino
s313439@studenti.polito.it

***Abstract*—Legal Named Entity Recognition (L-NER)[1] is considered crucial for the analysis of legal documents and the extraction of key information such as entities like court names, dates, and statutes. In this project, a specialized NER system is crafted for the legal field using advanced models such as LUKE and BERT. This system aims to identify and classify legal entities, aiding professionals in tasks like document analysis and information retrieval. The approach involved both training and inference phases, containing fine-tuning pre-trained models on legal datasets. The system is assessed using the NER evaluation metric, focusing on F1-score for performance. The inference phase saw the creation of an NER extractor to spot entities in unprocessed legal texts, incorporating trained models and tokenization. To enhance model performance and dataset diversity, we employed an extension. These included data augmentation through paraphrasing of less frequent label values to generate linguistically varied yet semantically consistent alternatives. Furthermore, we filtered out unlabeled data, removed HTML commands and punctuation, and focused on optimizing the dataset for key legal entities. Our system achieved promising results in legal NER tasks, demonstrating high F1-score and robustness in entity recognition. Through rigorous evaluation and extension methods, we improved model generalization and adaptability to diverse legal documents. The paraphrasing extension enriched the dataset, leading to more robust model training and better performance. Overall, our methodology and extensions contribute to advancing NER capabilities in the legal domain, offering valuable insights and practical applications for legal professionals.**

***Index Terms*—Legal Named Entity Recognition (NER), Dataset augmentation, F1-score evaluation metric**

I. Problem Statement

A. Expected Input

Each input element comprises the following elements, which is extracted form legal texts:

- Text Data: This component consists of unstructured legal text to undergo analysis. It encompasses legal documents where entities require identification and labeling.
- Annotations: Within the ”annotations” field resides an array containing labeled entity objects. Each entity object includes:
  - Starting and ending positions within the text.
  - Textual content of the labeled entity.
  - Assigned labels indicating its class (e.g., ”ORG” for organization, ”DATE” for date).

Fig.1 shows the initial number of entities for each label.

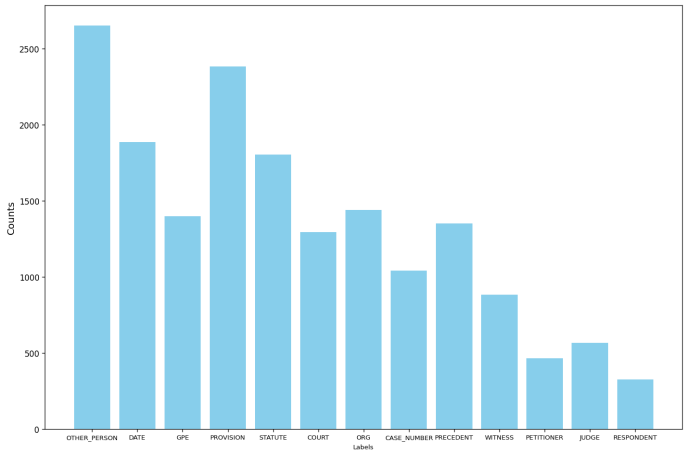

Fig. 1. Initial number of entities for each label

B. Addressed Task

The annotated samples, derived from legal documents, serve as training data for NER models specialized in the legal domain. These samples enable models to learn intricate patterns and associations within textual data, facilitating accurate recognition and categorization of various legal entities. During training, models iteratively adjust internal parameters based on annotated samples to enhance their ability to identify and classify legal entities effectively.

Following training, models undergo evaluation using a dedicated development dataset comprising labeled samples. This process is fundamental for assessing models' performance and generalization capabilities. By computing F1 scores, the evaluation phase provides insights into how accurately models identify legal entities across different legal contexts and document types. In the subsequent inference phase, trained models are deployed to actively extract named entities from unprocessed legal texts. This involves sophisticated tokenization techniques and model inference mechanisms to comprehensively identify and classify legal entities within the text. Through inference, models demonstrate practical utility in automating the extraction of legal entities, streamlining document analysis and information retrieval tasks for legal professionals.

These steps contribute to the overarching objective of developing and evaluating a specialized NER system tailored for the legal domain. By accurately identifying and classifying legal entities, the NER system enhances the efficiency and accuracy

of various legal processes, facilitating document analysis and information retrieval tasks.

C. Expected Output

Upon completion of the NER system development and evaluation process, expected outputs include several key components.

- Trained NER Models: These models accurately identify and classify legal entities within unstructured legal texts. They undergo iterative training processes using annotated samples to learn patterns and associations inherent in legal documents.
- Evaluation Metrics: The output comprises evaluation metrics, notably enhanced F1 scores, derived from the evaluation phase using a dedicated development dataset, offering insights into the models' performance and generalization capabilities. During the inference phase, deployed models actively extract named entities from unprocessed legal texts, showcasing their effective tokenization and inference abilities.
- Enhanced Efficiency: Ultimately, the output demonstrates the enhanced efficiency and accuracy of various legal processes facilitated by the developed NER system. By automating the extraction and classification of legal entities, the system streamlines document analysis workflows, enabling faster and more accurate information retrieval for legal practitioners and researchers.
- Contribution to the Objective: These outputs contribute to accurately identify and classify legal entities enhances the overall efficiency and effectiveness of legal document analysis and information retrieval tasks, benefiting the legal professionals.

II. Methodology

A. NLP Pipeline Overview

Our Natural Language Processing (NLP) pipeline consists of several key steps aimed at Named Entity Recognition (NER) within legal texts. The pipeline encompasses data preprocessing, model training, evaluation, and inference phases. Below is an overview of each step:

- Data Preprocessing:
  - Loading the dataset containing legal text samples and their corresponding annotations.
  - Cleaning the data by removing empty elements, HTML commands, and punctuation marks.
  - Adding extension, which is Augmenting the dataset through paraphrasing to improve model performance, particularly for less frequent labels (Fig.2 shows the number of entities for each label after extension.)[2]
- Model Training:
  - Utilizing pre-trained language models such as LUKE, BERT, and RoBERTa fine-tuned on legal domain-specific corpora.

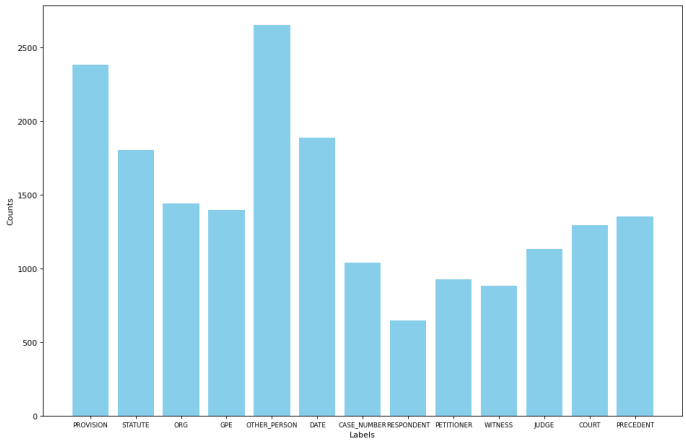

Fig. 2. Final Number of entities for each label

  - Training the models using the annotated dataset with labels for various entities including COURT, ORG, DATE, and others.
  - Employing token classification techniques to predict entity labels for each token in the input text.
- Evaluation:
  - Assessing the model's performance using metrics, which is F1-score.
  - Utilizing the nervaluate library to compute entity-level F1 scores for different evaluation strategies including type-match, strict, partial, and exact.
- Inference:
  - Employing trained models to perform NER on unseen legal text samples.
  - Extracting entities such as organizations, persons, dates, and other relevant information from the text.
  - Generating annotations with predicted entity spans and labels for further analysis or application.

B. Modules Description

- NERExtractor:
  - A class responsible for extracting named entities from text using pre-trained token classification models.
  - Utilizes tokenizer and model checkpoints to process input text and generate predictions for entity labels.
  - Implements methods for tokenizing text, generating predictions, and aligning entity labels with tokenized input
- LegalNERTokenDataset
  - A dataset class tailored for NER tasks in the legal domain.
  - Loads legal text samples and their annotations from JSON files.
  - Tokenizes text using appropriate tokenizers (e.g., RoBERTaTokenizerFast, AutoTokenizer).

  - Matches annotations with tokenized input to generate aligned entity labels for training and evaluation.
- Data Cleaning
  - Initial Cleaning: The dataset is first processed to remove elements where the annotated (labeled) results are empty, ensuring that only text samples with relevant annotations are included for training and evaluation.
  - Removing Noise: Subsequently, the text data undergoes further cleaning to eliminate HTML tags and punctuation marks. This step is crucial for normalizing the text and ensuring that the model's focus remains on meaningful content. Adjustments are also made to the start and end indices of labeled entities to account for the removal of these characters, maintaining the accuracy of annotations.
- Data Filtering for Extension(Augmentation)
  - Identification of Less Frequent Labels: The cleaned dataset is then analyzed to identify entities with less frequent labels such as 'JUDGE', 'RESPONDENT', and 'PETITIONER'. These labels are targeted for augmentation to address data imbalance and enhance the model's ability to recognize these entities accurately.
  - Paraphrasing for Augmentation: Texts containing these less frequent labels undergo paraphrasing, a process designed to generate textual variations without altering the semantic content. The paraphrasing module leverages NLTK and WordNet to find synonyms and replace words according to their part-of-speech tags, ensuring that the paraphrased text remains coherent and relevant to the original context.
- Integration with Main Dataset
  - Merging: The paraphrased texts are then integrated back into the main dataset. This step enriches the dataset with additional variations for the targeted less frequent labels, aiming to balance the dataset and improve the model's exposure to a wider range of expressions and usages of these entities.
  - Enhanced Training Corpus: The result is a more balanced and diversified training corpus that includes a higher representation of less frequent entity labels. This augmentation process is expected to boost the NER model's performance by improving its generalization capabilities across a broader spectrum of legal text samples.

This methodology ensures a comprehensive approach to NER in the legal domain, integrating data preprocessing, model training, evaluation, and data augmentation techniques to enhance model performance and generalization capabilities.

III. Experiments

A. Data Description

The dataset used in the experiments consists of legal text samples annotated with labeled entities relevant to the legal domain. Each sample includes unstructured text data along with annotations specifying the starting and ending positions of labeled entities, along with their corresponding labels (e.g., "ORG" for organization, "DATE" for date).

B. Experimental Design

- Hardware: The experiments were conducted on Google Colabratory with NVIDIA Tesla T4 GPU.
- Software Libraries: The experiments utilized the following software libraries:
  - PyTorch version 1.13.1 for training and inference of deep learning models.
  - Transformers version 4.36.0 library for utilizing pretrained language models like BERT, RoBERTa, and LUKE.
  - Scikit-learn version 1.2.1 for evaluating model performance using metrics like F1-score.
  - NLTK for text processing, tokenization, and paraphrasing.
  - BeautifulSoup for removing HTML commands from texts.
  - Matplotlib for plotting the number of values in each label and comparison.

C. Validation Method and Performance metrics

The performance of the models was evaluated using entity-level F1 scores computed using the nervaluate library. Different evaluation strategies including type-match, strict, partial, and exact were employed to assess the models' performance comprehensively. The primary performance metric used in the experiments is the F1-score, which provides a balanced measure of a model's precision and recall in identifying legal entities.

D. Execution Times

The evaluation steps per second for the experiments, encompassing model training, evaluation, and inference, were meticulously recorded. Due to the computational complexity involved in training large language models and performing inference on legal text datasets, the LUKE Large model took longer to process than others, reflecting the intensive computational demands of these operations.

IV. Results

Table1 illustrates the initial and final F1 scores for the models, while Table2 delineates the runtime of the respective models.

A. Performance Overview

- Top Performers: The final version of Luke-base stands out with the highest F1 scores across all metrics, achieving an F1-strict score of 81.20%, F1-partial of 87.27%, F1-exact of 83.10%, and F1-type match of 87.54%. This indicates its superior ability in precisely identifying and classifying named entities in legal texts.

TABLE I
F1 SCORES FOR DIFFERENT MODELS

Models	strict	partial	exact	TypeMatch
LegalBERT-base (I)	78.54%	85.80%	80.59%	86.10%
LegalBERT-base (F)	80.43%	86.87%	82.40%	86.80%
LegalRoBERTa-base (I)	79.12%	86.75%	81.34%	86.98%
LegalRoBERTa-base (F)	77.19%	83.64%	78.61%	84.96%
EURLEX (I)	77.14%	84.81%	78.83%	85.88%
EURLEX (F)	78.97%	85.71%	80.84%	86.18%
BERT-ECHR (I)	75.82%	84.55%	78.52%	83.61%
BERT-ECHR (F)	77.96%	85.37%	79.75%	86.35%
Luke-base (I)	79.59%	86.41%	81.13%	86.66%
Luke-base (F)	81.20%	87.27%	83.10%	87.54%
BERT-large-NER (I)	0%	0%	0%	0%
BERT-large-NER (F)	0%	0%	0%	0%
Luke-large (I)	0%	0%	0%	0%
Luke-large (F)	0%	0%	0%	0%

TABLE II
RUNTIMES FOR DIFFERENT MODELS

Models	Runtime
LegalBERT-base (initial)	01:19:59
LegalBERT-base (final)	01:23:21
LegalRoBERTa-base (initial)	02:56:49
LegalRoBERTa-base (final)	02:59:41
BERT-base-EURLEX (initial)	01:20:46
BERT-base-EURLEX (final)	01:27:03%
BERT-ECHR (initial)	01:17:44
BERT-ECHR (final)	01:25:53
Luke-base (initial)	02:50:15
Luke-base (final)	02:48:27
BERT-large-NER (initial)	02:10:02
BERT-large-NER (final)	02:18:14
Luke-large (initial)	08:03:07
Luke-large (final)	8:50:46

- Improvement Over Iteration: Most models show an improvement from their initial to final versions, suggesting that further training or fine-tuning on legal texts enhances their NER capabilities. Notably, LegalBERT-base and BERT-base-EURLEX show consistent improvements across all F1 metrics, highlighting the effectiveness of domain-specific pre-training and fine-tuning.
- Underperformers: BERT-large-NER and Luke-large, performed poorly with 0% F1 scores, indicating potential issues with model configuration or compatibility with legal NER tasks. LegalRoBERTa-base showed declining F1 scores in its final version, suggesting problems like overfitting or inadequate adaptation to legal text NER requirements. These findings emphasize the need for nuanced model optimization, especially in domain-specific applications like legal NER.

B. Runtime

- Efficiency:LegalBERT-base, in its final form, offers a good balance between high F1 scores and reasonable runtime, taking 01:23:21, which is efficient compared to the more accurate but significantly slower Luke-base (final) model.
- Longest Runtimes: Luke-large models have the longest runtimes, with the final version reaching almost 9 hours. Despite their potentially advanced capabilities due to their size, their performance metrics are at 0%, which, combined with their long runtime, suggests there might be a problem with how these models are being applied to the task at hand.
- Comparative Efficiency: The runtime increase from initial to final versions is generally modest across models that show performance improvement, indicating that additional training time is relatively well-compensated by the gains in accuracy.

C. Analysis of Results

- Model Optimization: The analysis indicates a discernible correlation between model optimization efforts, such as further training or fine-tuning, and improved performance in legal Named Entity Recognition (NER) tasks. However, it is noteworthy that while most models demonstrate enhanced performance with additional iterations, the final version of LegalRoBERTa-base presents a deviation, experiencing a decrease in performance. This anomaly suggests potential challenges such as overfitting or suboptimal training strategies that may hinder the efficacy of certain models despite optimization attempts.
- Model Selection:
  Luke-base (final) emerges as the most accurate model for legal NER tasks based on achieved F1 scores. However, its runtime may limit its suitability for certain applications. Conversely, LegalBERT-base (final) offers a favorable combination of high accuracy and manageable runtime, making it a well-rounded choice for various practical scenarios.
- Large Models' Anomaly: The absence of performance in large models like BERT-large-NER and Luke-large despite their theoretically greater learning capacity raises critical questions. Potential issues may include data preprocessing errors, model configuration discrepancies, or a fundamental misalignment between the models' architectures and the specific nuances of legal text, independent of any fine-tuning or optimization efforts. However, in the context of legal NER, evaluation primarily hinges on the achieved F1 scores, regardless of runtime considerations.

D. Conclusions and Recommendations

This analysis underscores the importance of model selection based on both accuracy and efficiency for legal NER tasks. While the Luke-base model in its final iteration shows the best performance in terms of F1 scores, it is practicality limited by long runtimes. LegalBERT-base (final) offers a compelling alternative with its balance between performance and efficiency. The failure of larger models underscores potential challenges within the project, rather than suggesting a blanket assertion that bigger models are inherently inferior. Future work should explore the reasons behind the underperformance of larger models and continue to refine those with promising results, like LegalBERT and Luke-base, to enhance both their accuracy and efficiency.

REFERENCES

[1] PoliToHFI at SemEval-2023 Task 6: Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Elena Baralis, Luca Cagliero, Francesco Tarasconi

[2] Wei, Jason, et al. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." arXiv preprint arXiv:1901.11196 (2019).