Legal Named Entity Recognition

Hesam Khanjani
Politecnico di Torino
Student ID: 310141
s310141@studenti.polito.it

Mohamad Samaei
Politecnico di Torino
Student ID: 314577
s314577@studenti.polito.it

Wiktoria Woronko
Politecnico di Torino
Student ID: 320150
s320150@studenti.polito.it

Sepehr Alemzadeh
Politecnico di Torino
Student ID: 314315
s314315@studenti.polito.it

***Abstract*—Named Entity Recognition (NER) is a fundamental Natural Language Processing (NLP) technique that entails identifying and categorizing entities, including names of individuals, organizations, locations, dates, and other specific terms, within a given text. This paper addresses Legal Named Entity Recognition by exploring two different methods aiming at improving accuracy of the model and addressing the limitations of legal NER. The first solution leverages Data Augmentation techniques. The experimental results show an improvement in performance compared to the baseline. The second solution introduces a completely new model using Conditional Random Fields (CRF) that lays the foundation for future works as well as giving a comparable performance without requiring strong computational resources and training time.** [1]

I. Problem Statement

A. Expected Input

1) Datasets: The project enjoys two datasets: one to train the model and another in order to test the performance. The training dataset is split by the percentage of 80 to 20 to create another dataset for validation. Hence, in the train dataset we have 7548 sentences and in the validation dataset we have 1887 sentences. The same applies to the test dataset having 949 number of sentences. In all the datasets, we have 13 entity classes comprised of case number, court, date, GPE, judge, organisation, petitioner,precedent, provision, respondent, statue, witness and other person.

2) Models: In the first contribution of our work we make use of BERT [1] and RoBERTa generically fine-tuned for the NER task (BERT-large and RoBERTa-large respectively). In addition to this, we have our BERT and RoBERTa [5] model specifically fine-tuned for the legal domain: Legal BERT-base and Legal RoBERTa-base, EURLEX (BERT-base), and ECHR (BERT-base).LUKE (Language Understanding with Knowledge-based Embeddings) is a transformer-based architecture that incorporates entity information into pre-trained language models. [9]

B. Addressed Task

In Named Entity Recognition the main issue is labeling each word in the text with a certain class, a process called Sequence Labeling in this domain. In our work, these labels are legal-domain-related meaning that labels such as ”COURT” or ”CASE NUMBER” are seen abundantly. A detailed list of entity names can be seen in Figure 1. Indeed, what makes this project challenging is having an application in the domain of law (L-NER). In this work we introduce two contributions: First is augmenting the data with a modification of the approach mention replacement. The baseline is implemented using transformer-based models including BERT, RoBERTa, and LUKE. The second contribution is looking at the problem from a different angle. Enjoying the same dataset, we approach the problem via a probabilistic model referred to as CRF.

[1]Git repository github.com/woronkowiktoria/Legal-Named-Entity-Recognition.

C. Expected Output

Table I and Table II indicate the details of default approach and modifications done by our team. The first one is related to the first contribution and the second one related to the second contribution.

II. Related works

This paper is an extension of an ablation study on transformer-based models for Legal NER [4]. The dataset was provided by the authors of [7] L-NER baseline model.

A. Data Augmentation

The Syntax-driven Data Augmentation for Named Entity Recognition delves into the efficacy of data augmentation strategies with low-resource scenarios. An innovative augmentation approach employing constituency-tree mutations [8], designed to preserve linguistic cohesion in augmented sentences is proposed. A comparative analysis is conducted with conventional techniques such as masked language model replacement, as well as established strategies like synonym-based and mention-based replacement. The experimentation utilizes the i2b2-2010 dataset, revealing that while all augmentation methods contribute to improved NER performance, certain approaches may exhibit diminished effectiveness or even prove detrimental as the size of the training data increases.

The effectiveness of elementary augmentation strategies, including label-wise token replacement, synonym replacement, mention replacement and segment shuffling [3] has been shown in the results, which indicates that these uncomplicated augmentation techniques play a pivotal role in enhancing NER performance, particularly in scenarios with limited training data in domain-specific contexts.

B. Conditional Random Fields

A statistical Named Entity Recognition system leveraging machine learning to detect and categorize named entities in

Marathi language [6] texts is proposed. This system utilizes Conditional Random Fields (CRF) for the identification and classification of Named Entities (NE). Statistical algorithms achieve considerable accuracy in NE identification and classification, but they require additional knowledge for further accuracy improvements. Inspired by the use of this discriminative model, in this paper we explore the CRF model with Indian legal dataset in the English language.

III. Methodology

A. Data augmentation

In Legal NER, a specific application in the legal domain, each entity is related to a particular task. In the beginning, we tried using the Spacy library but we did not achieve a satisfying outcome because the library performed the augmentation in a generic manner and most importantly it did not augment the data professionally since it did not respect the positioning of each entity. The problem of positioning arises from the fact that Spacy replaces named entities of different length with each other. For example, it would replace "Hongkong Bank" with "Rahul Co.". Although they are both from the entity type "ORG", they are of different lengths. Hence, using Spacy would disturb the positioning in the dataset.

Simultaneously, we know for a fact that improving annotations require expert knowledge of law in India, therefore we implemented a method called Mention Replacement which could automatically enhance the size of the dataset by using the data inside the dataset. By creating a dictionary derived from the entire training dataset, we facilitated the categorization of entities based on shared characteristics, particularly focusing on their length. Subsequently, our augmentation technique involved comparing a specific entity with the pre-established dictionary to identify components with similar lengths. To ensure diversity in replacements, a random selection was made in those dictionary components that shared the same entity type and length. This mitigates potential errors and enhances the robustness of the model. [2] As a result of this our approach, the accuracy improved around 2%.

The applicability of this approach is underscored by the unique characteristics of the Legal Domain, where entities such as names and organizations correspond intricately to specific countries. Leveraging an Indian dataset, which aligns with the Legal domain and its associated terms, proved crucial for adaptability. Therefore, we utilized the Indian dataset itself; this tailored approach not only addresses the contextual relevance of the dataset, but also contributes to the overall efficacy of the mention replacement technique within the Legal NER framework.

B. Conditional Random Fields

CRF is a type of discriminative model, used to calculate the conditional probability of hidden states in the model with respect to observed states. It was chosen for its flexibility when it comes to the selection of feature functions needed for the training. Figure 1 well depicts its pipeline. The architecture is described below.

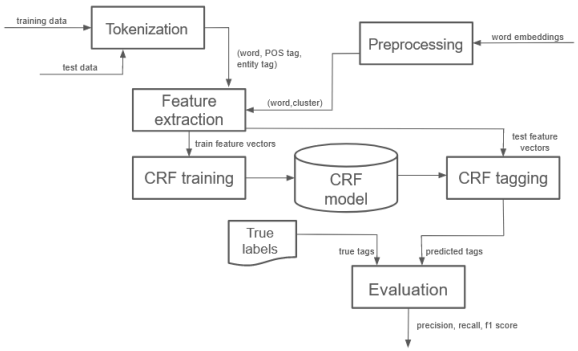

Fig. 1. CRF system architecture.

Data preparation All the datasets used for training, validation and testing were tokenized with tagged sentences. Each token is in the format ('word', 'POS tag', 'entity tag') which follows BIO convention, that will be later used for feature extraction.

Feature extraction

In the feature extraction phase for Conditional Random Fields (CRF), each word is converted into a feature vector. This vector encompasses the word itself and its respective part-of-speech tag, along with indicators for capitalization and digit presence. Additionally, character-level bigrams and trigrams located at the end of the word are incorporated, alongside a bias term heloping in the understanding of label frequency relative to training data. To give the CRF more information about the meaning of a word, embeddings are also used. They are deriven from a Word2Vec model, taken from SigmaLaw - Large Legal Text Corpus and Word Embeddings[2], which was built on US Supreme Court legal cases. The word embeddings were then clustered into 500 groups based on legal text data. These embeddings augment the semantic understanding of words. Each word is then mapped to the id of the cluster it is in. This is useful for NER, as most entity types cluster together and allow CRFs to generalize above the word level.

Training The standard L-BFGS algorithm is used for parameter estimation and run for 100 iterations.

Fig.1 fully presents the model architecture.

IV. Experiments

A. Data description

The dataset contains 13 named entities classes, which are case number, court, date, GPE, judge, organisation, petitioner, precedent, provision, respondent, statue, witness and other person. For both extensions, the train dataset corpus is composed of 9435 sentences and was split into training and validation datasets in 80:20 proportion. Dev dataset has 949 sentences, and was used for testing. Figure 2 presents the number of named entity instances present in the test corpus.

[2]https://osf.io/qvg8s/

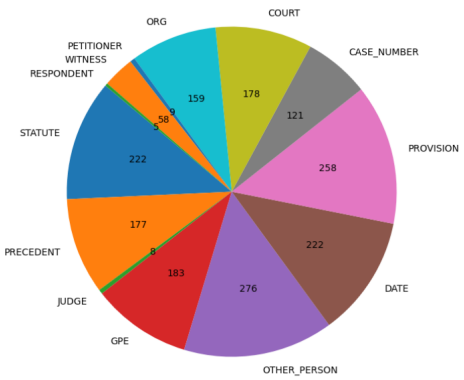

Fig. 2. Number of named entity instances in the test corpus.

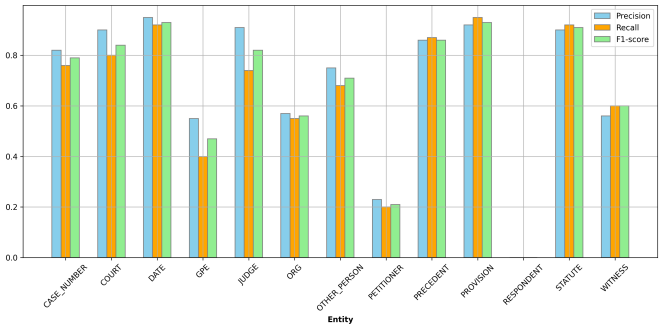

Fig. 3. F1 score, precision and recall comparison for each named entity class in the test dataset.

B. Experimental design

Libraries For the CRF, sklearn-crfsuite was used. It's a wrapper around python-crfsuite, which itself is a Python binding of CRFSuite.

Evaluation For testing both extensions, the models are evaluated by using strict, partial, exact, and type-match F1 scores on the judgment sentences. In the CRF model specifically, standard F1 score, standard precision and recall, was chosen for per entity evaluation computed with Named-Entity Recognition evaluation metrics[3]. The details of per entity evaluation can be seen in Figure 3

C. Results and analysis of the results

*1) **Data Augmentation**:*
According to Table I, the **EURlex** model, when augmented with the WITNESS extension, demonstrates superior performance within its designated group across all measured metrics. While the amalgamation of JUDGE + COURT outperforms the Default configuration, it falls short of achieving the level of effectiveness exhibited by the WITNESS extension.

In the evaluation of the **legal-bert-base-uncased** model, a significant enhancement is observed upon the implementation of the JUDGE modification, surpassing the default setting across all assessed metrics. However, the introduction of the JUDGE + COURT adjustment does not yield substantial improvement compared to the Default setting, with performance levels remaining nearly equivalent between the two configurations.

Within the context of the **legal-roberta-base** model, the utilization of the WITNESS extension stands out as the most effective, showcasing superior performance across all evaluated metrics and surpassing alternative configurations. Although the JUDGE + COURT extension exhibits an enhancement in the F1 Partial score, it does not surpass the effectiveness of the WITNESS modification. Furthermore, the JUDGE extension demonstrates a comparatively weaker performance in F1 Type Match when compared to the Default setting. Consequently, the most optimal choice based on comprehensive evaluation is the WITNESS extension.

In the evaluation of the **bert-base-uncased-echr**, it is observed that the extension JUDGE yields the most optimal performance across all metrics. Additionally, the combined use of JUDGE + COURT exhibits a marginal improvement in two metrics, specifically F1 exact and F1 partial, when compared to the default setting. However, on a holistic assessment, the overall performance of the combined extension is deemed comparable to that of the default configuration. Consequently, based on the comprehensive evaluation, it is inferred that the most advantageous extension for optimizing the model's performance is JUDGE.

Finally, for **studio-ousia/luke-base** model, the default configuration exhibits commendable F1 Partial and F1 Type Match scores, showcasing competitiveness within its performance metrics. Introducing the JUDGE modification, however, does not notably affect the F1 Exact and F1 Strict scores but leads to a slight reduction in F1 Partial and F1 Type Match scores. Furthermore, the addition of JUDGE + COURT yields a less favorable outcome in both F1 Exact and F1 Partial metrics. Thus, on an overall evaluation, the JUDGE configuration proves to be the most optimal choice for this model.

*2) **Conditional Random Fields**:* We had insignificant data sources for legal Indian. As it is shown in Figure 3, some of the entities, e.g. petitioner or respondent, have resulted in poor performance metrics because the word embedding used in the training were based on the US legal cases and not on the Indian legal cases. Moreover, the support for those entities in the test corpus was significantly smaller than for other classes present in Figure 2. Conversely, other entities, e.g. court or date, which were also more frequent in the training dataset, showed very good performance of over 80% f1 score.

Table II shows the overall comparison between the baseline models and the CRF. The CRF scored 78.23% for F1 exact score, 74.07% for f1 strict, 82.84% for f1 partial and 80.34% for f1 type match score. Overall, the F1 scores of CRF are lower with respect to the transformer-based models, however, the difference is in the range of 4-8%.

[3]https://github.com/davidsbatista/NER-Evaluation

Model	Change	F1 Exact	F1 Strict	F1 Partial	F1 Type Match
eurlex	Default	0.8130	0.7907	0.8625	0.8601
eurlex	JUDGE	0.8124	0.7996	0.8640	0.8748
eurlex	WITNESS	0.8234	0.8055	0.8730	0.8780
eurlex	JUDGE + COURT	0.8263	0.7963	0.8777	0.8707
eurlex	WITNESS + JUDGE + COURT	0.82301	0.8019	0.8757	0.8754
legal-bert-base-uncased	Default	0.8310	0.8079	0.8763	0.8745
legal-bert-base-uncased	JUDGE	0.8403	0.8202	0.8830	0.8814
legal-bert-base-uncased	JUDGE + COURT	0.8346	0.8079	0.8803	0.8792
legal-roberta-base	Default	0.8303	0.8136	0.8782	0.8858
legal-roberta-base	JUDGE	0.8340	0.8158	0.8795	0.8847
legal-roberta-base	WITNESS	0.8405	0.8197	0.8848	0.8865
legal-roberta-base	JUDGE + COURT	0.8369	0.8182	0.8822	0.8858
bert-base-uncased-echr	Default	0.8154	0.8006	0.8677	0.8796
bert-base-uncased-echr	JUDGE	0.8268	0.8078	0.8764	0.8809
bert-base-uncased-echr	JUDGE + COURT	0.8232	0.8068	0.8729	0.8779
studio-ousia/luke-base	Default	0.8294	0.8097	0.8779	0.8809
studio-ousia/luke-base	JUDGE	0.8363	0.8184	0.8767	0.8801
studio-ousia/luke-base	JUDGE + COURT	0.8273	0.8127	0.8750	0.8845

TABLE I
COMPARISON BETWEEN MODELS WITH DIFFERENT EXTENSIONS

Model	F1 Exact	F1 Strict	F1 Partial	F1 Type Match
eurlex	0.8130	0.7907	0.8625	0.8601
legal-bert-base-uncased	0.8310	0.8079	0.8763	0.8745
legal-roberta-base	0.8303	0.8136	0.8782	0.8858
bert-base-uncased-echr	0.8154	0.8006	0.8677	0.8796
ECHR	0.8188	0.7995	0.8711	0.8864
studio-ousia/luke-base	0.8294	0.8097	0.8779	0.8809
CRF	0.7823	0.7407	0.8284	0.8034

TABLE II
RESULTS OF THE CRF MODEL WITH RESPECT TO TRANSFORMER-BASED MODELS.

V. CONCLUSION

In the conclusion of this study, we aimed to address the inherent limitations of Named Entity Recognition (NER) in the complex domain of legal texts. Legal language presents unique challenges due to its specialized vocabulary and intricate sentence structures. To overcome these, we implemented two distinct but complementary approaches: Conditional Random Fields (CRF) and data augmentation.

Also, Data Augmentation in the domain of law and another language requires knowledge of the culture and language as well as the knowing the law itself of the country. For These reasons, Augmentation would be a challenging task to handle. So, to address this issue we implemented that modification to mention replacement. As a result as the dataset grow the model is better able to gather more entities of the same type and of the same length and replace them. so this is how could improve the task of NER in law without knowing the language and law. If in the future works we have access to a big dataset, we should consider that overfitting might occur.

A. Remarkable Outcomes

The CRF brought a level of precision by leveraging contextual clues within the text, thus improving the granularity with which entities were recognized. Data augmentation, on the other hand, allowed us to expand our dataset significantly, providing our models with a broader spectrum of examples to learn from. Together, these methodologies not only enhanced the model's ability to understand and categorize legal terminology but also ensured robustness against the variability inherent in legal documents.

Within the scope of Augmentation, to address this issue we implemented a modification to the approach of mention replacement. As a result, as the dataset grows the model is better able to gather more entities of the same type and length and replace them with each other. Thus, this is how could improve the task of NER in law without knowing the language and law.

B. Main Challenges

Having few available sources for legal Indian data presented a challenge since some of the entities, such as petitioner or respondent, had their word embedding trained based on the US legal cases. Another major challenge in the work was the domain of legal text which is concomitant with vocabulary of the legal nature.

Data Augmentation requires knowledge of the culture and language as well as knowledge of how that specific domain is practiced in the respected country (in our work the domain was law). For These reasons, Augmentation would be a challenging task to handle.

C. Lesson Learned

The performance of the CRF substantiates that without complex transformer-based architectures, we could achieve approximately as much performance using Conditional Random Fields. In addition to this, achieving a better performance in the task of Named Entity Recognition using is about understanding the data in order to augment them in a way that improves the performance significantly. Finally, the synergy between contextual sensitivity of the CRF and the enriched diversity from data augmentation marked a substantial advancement in legal NER capabilities.

In the case of Augmentation, It worth noting that in the future works having access to a bulky dataset does not suffice since methods of addressing overfitting should be implemented simultaneously.

REFERENCES

[1] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

[2] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. Data augmentation for cross-domain named entity recognition. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683*, 2020.

[4] Lorenzo Vaiani Eliana Pastor Elena Baralis Luca Cagliero irene Benedetto, Alkis Koudounas. Politohfi at semeval-2023 task 6: Leveraging entity-aware and hierarchical transformers for legal entity recognition and court judgment prediction. *Association for Computational Linguistics*, 2023.

[5] Yinhan Liu, Myle Ott, and Naman Goyal. Jingfei du, mandar joshi, danqi chen, et al.,"roberta: A robustly optimized bert pretraining approach,". *arXiv preprint arXiv:1907.11692*, 2019.

[6] Nita Patil, Ajay Patil, and BV Pawar. Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188, 2020.

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

[8] Arie Pratama Sutiono and Gus Hahn-Powell. Syntax-driven data augmentation for named entity recognition, 2022.

[9] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020.