Towards Legal-NER Across Languages: Multilingual Framework and Evaluation

Pablo Borrelli
Politecnico di Torino
s303878@studenti.polito.it

Daniele Mansillo
Politecnico di Torino
s319297@studenti.polito.it

Umberto Picone
Politecnico di Torino
s296496@studenti.polito.it

Luca Varriale
Politecnico di Torino
s300795@studenti.polito.it

ABSTRACT

Legal Named Entity Recognition (L-NER) poses unique challenges due to the intricate legal language and varied structures of legal documents. Adapting NER models to the legal context is additionally complicated in the multi-language domain, considering the diverse legal landscapes across countries. This study employs dedicated models like various Legal BERTs and LUKEs, both tailored for monolingual and multilingual tasks, along with a novel multilingual dataset, to develop a comprehensive NER model for analyzing legal documents across languages. The research evaluates model performance on both language-specific datasets, explores the intricacies of model embeddings using t-SNE visualizations, and assesses the effectiveness of multilingual adaptation. The results indicate a promising performance from multilingual LUKE, especially on a unified dataset, showcasing its adaptability to diverse legal contexts. The t-SNE visualizations reveal distinct patterns in entity categorization among different models, offering insights into their interpretability and performance nuances. This research contributes with useful insights for advancing NER models in the legal domain, emphasizing the importance of datasets and labels in multilingual adaptation.
The associated code and dataset, available for reference and collaboration, can be accessed through the provided repository at GitHub.

I. Problem Statement

Named Entity Recognition is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories, in our case these categories are strictly related to the legal domain.

Legal documents, characterized by a lexicon full of nuanced language and complex structures, pose an arduous challenge to conventional NER models designed for general use. The distinctive traits of legal language, including specialized vocabulary, intricate acronyms, different document structures and the prevalence of specific entity types, necessitate a tailored approach to NER in the legal domain.

Adapting a NER model to a multi-language domain introduces additional complexities. The legal landscape varies across countries, and configuring a model to function effectively not only across diverse languages but also distinct legal systems proves to be a challenging task. Moreover, the dynamic nature of legal systems introduces fluctuations in vocabulary, diverse acronyms, and varying document structures, presenting significant obstacles in the adaptation process.

By employing dedicated models designed for multi-language tasks, such as multilingual LUKE, along with a novel multilingual dataset, our aim is to develop a comprehensive model proficient in seamlessly analyzing legal documents across diverse languages.

To achieve this goal we first identify some performance baseline over different language datasets and subsequently fine-tune the model on a multilingual dataset to assess if the performance remain satisfying also in a multilingual domain.

In our journey to understand how well our models work, we took a closer look at the Named Entity Recognition (NER) module. We wanted to analyze what goes on inside the model when it tries to recognize different types of named entities. Employing the dimensionality reduction technique known as t-SNE, we translated the complex model embeddings into visual representations. By comparing these visualizations, our aim was to discern subtle divergences in how each model interprets and categorizes entities such as names, dates, and locations. This exploration helped us uncover unique patterns and details in the way each model handles this important task.

II. Methodology

In this study, we leverage multilingual extensions of NER systems tailored to the legal domain in order to enhance the language capabilities of the model. The analysis stems from the research of publicly available datasets in different languages, evaluating their accessibility and suitability concerning the legal NER task addressed in [8]. Specifically, languages approached in the following analysis are English, German and Spanish.
At first, we conducted separate experiments on the Spanish

and German datasets training two different models, tailored to each language domain. The results of these and subsequent analyses are reported in the following section.

Then, we created a multilingual dataset by merging the English, Spanish, and German datasets. The cross-language training and test on the unified dataset were conducted using a LUKE model [13], that is based on an entity-aware attention mechanism thanks to which it generates contextualized representations for both words and entities simultaneously. For the merging of datasets, label alignment was conducted to guarantee consistent evaluation across languages.

An important focus during the comparative analysis is to evaluate the quality of the produced embeddings. Specifically, after extracting embeddings from different models addressing the L-NER task, a visual representation of these complex data can improve the interpretability of models utilized. In this regard, the t-Distributed Stochastic Neighbor Embedding (t-SNE), a statistical method for visualizing high-dimensional data, is a powerful tool that has been employed to gain insight into entities distribution, which reflects semantic relationships among them.

After reproducing the performance results of various models on a monolingual dataset [8], the embeddings generated by each model were extracted to provide a two-dimensional visualization in order to enhance the interpretability of the results and visualize the generated embeddings of the models.

III. Experiments

A. Datasets

Legal document collections are characterized by intricate and formal language and the entity types relevant for NER tasks are highly domain-specific. Moreover legal texts may be subject to rigorous confidentiality and privacy regulations, thus influencing the limited availability of annotated data. In our study, we employed datasets related to three distinct language-specific legal corpora, respectively in English, German, and Spanish.

- English dataset [9], provided by OpenNyAI, is an open-source dataset developed to implement AI-powered solutions to enhance access to justice in India. It includes following named entities for L-NER task: *court, petitioner, respondent, judge, lawyer, date, org, gpe, statute, provision, precedent, case number, witness* and *other person*.
- German dataset [10] was developed for NER tasks in German federal court decisions and it includes approximately 67,000 sentences with 54,000 manually annotated entities. The entities are mapped into 19 fine-grained semantic classes and all of which were utilized during preliminary analyses. However, for the purpose of label alignment, as regards multilingual application all these labeled entities have been grouped into 7 broader labels, including all the typical NER classes and some semantically specific to the legal domain. Their number and their percentage compared to the total number of entities are reported in Table I. The German dataset was provided in a format which was not compatible with the python project we were working on, this required the development of a specific script to extract the needed information, adapt the format of the German dataset and save it in a suitable format.

	Classes	#	%
PER	Person	3,377	6.30
LOC	Location	2,468	4.60
ORG	Organization	7,915	14.76
NRM	Legal norm	20,816	38.81
REG	Case-by-c. regul.	3,470	6.47
RS	Court decision	12,580	23.46
LIT	Legal literature	3,006	5.60

TABLE I: Distribution of entity classes in the German dataset.

- Spanish dataset [11] constitutes a subset derived from the meta-corpus Legal-ES [12]. This specific subset has undergone annotation with the explicit objective of facilitating the training of Named Entity Recognition models tailored to the domain of legal documents. Due to the focus on the legal domain, the texts comprising this dataset are sourced from the *BOE-A* and *DOUE* categories within the Legal-ES dataset. The former category encompasses a compilation of Spanish national laws, while the latter comprises spanish written legislation from the European Union. The number and percentage of the annotated entities [1] are reported in Table II.

	Classes	#	%
PER	Person	2204	12.80
LOC	Location	1367	7.94
ORG	Organization	6429	37.33
TIME	Time	1731	10.05
LEGAL	Legal references	5493	31.89

TABLE II: Distribution of entity classes in Legal-ES.

- A novel multilingual dataset targetted at Legal-NER tasks composed of a combination of sentences in English, German and Spanish extracted from the aforementioned datasets.
  To construct the dataset, we initially aligned the labels across all three datasets. Limited by the fact that the Spanish dataset had the fewest labels, we adjusted the labels in the other two datasets accordingly by grouping togheter multiple labels under a unique mapping as outlined in Tables III and IV. Subsequently, we balanced the number of observations by sampling an equal number from each dataset, adjusting the sample size to match the dataset with the fewest sentences, being once again the Spanish dataset, which had 6783 sentences for the training set and 798 for the validation set. In the end, we obtained a training dataset comprising 20,349 sentences and a validation dataset comprising 2394 sentences. We

[1]The complete dataset is comprised of more than 10M annotated named entities but for our purposed only the data publicly avaliable at *https://github.com/dosamy/NERC-Legal-Es-Example-* was used

opted to keep separate the test datasets to conduct more comprehensive tests on the model's performance with specific languages.

English	Multilingual
COURT	LEGAL
PETITIONER	PER
RESPONDENT	PER
JUDGE	PER
DATE	TIME
ORG	ORG
GPE	LOC
STATUTE	LEGAL
PROVISION	LEGAL
PRECEDENT	LEGAL
CASE_NUMBER	LEGAL
WITNESS	PER
OTHER_PERSON	PER
LAWYER	PER

TABLE III: Mapping for the entity classes in the English dataset to the multilingual.

German	Multilingual
LIT	LEGAL
LOC	LOC
NRM	LEGAL
ORG	ORG
PER	PER
REG	LEGAL
RS	LEGAL

TABLE IV: Mapping for the entity classes in the German dataset to the multilingual.

B. Experimental Setup

The training of the Legal NER models was carried out in Google Colab© using a 16GB V100 GPU. In order to match the resources offered by Google Colab© we used a subset of the total sentences when training the multilingual models to accommodate the constraints imposed by computational resources. We were limited to a batch size of 32 due to the lack of additional VRAM. Since our goal was mainly comparing the performance of different models over Legal-NER tasks, to maintain consistency across all experiments and to obtain comparable results, we applied a uniform set of hyperparameters which are reported in table V.

Hyperparameter	Value
batch size	32
epoch	5
learning rate	1e-4
weight decay	0.01
warm-up ratio	0.06

TABLE V: Hyperparameters used across all experiments.

To ensure repeatability of the experiments, they where all performed inside of a Python virtual environment running Python 3.10 and installing only the libraries indicated in the *requirements.txt* file.

In the individual experiments carried out on the Spanish and German datasets, we utilized two distinct models customized for each language domain. For the German dataset we employed a fine-tuned version of BERT-base-german-cased [2], while the model trained on Legal-ES dataset is PlanTL-GOB-ES/roberta-base-bne-capitel-ner [5], a NER model for the Spanish language fine-tuned from the roberta-base-bne model, a RoBERTa base model pre-trained using a large Spanish corpus compesed by the National Library of Spain (Biblioteca Nacional de España). In the multi-language experiments we utilized a general multilingual model: studio-ousia/mluke-base, training it each time on the different dataset we wanted to analyze.

C. Performance metrics

The evaluation metric employed during our analyses is the F1 score, which combines precision and recall to provide a comprehensive measure of the model's performance. Specifically, we included strict, partial, exact, and type-match F1 scores, which present various constraints between the predicted entity and the gold standard, in terms of matching between entity types and the correspondence of surface string:

- *Strict*: perfect matching between their entity types and their surface string boundaries.
- *Exact*: exact matching between entities' surface string boundaries, regardless of their entity types.
- *Partial*: it occurs when the predicted entity boundaries partially align with the gold standard entity, covering only a portion of their boundaries, regardless of entity types.
- *Type-match*: requires some overlap between the predicted and gold standard entities, along with a match in their entity types.

D. Results

Table VI displays the outcomes of the training sessions on language-specific datasets. The first set of results corresponds to the use of language-specific NER models, followed by the outcomes when employing a multilingual LUKE [7].

Looking at the results from the individual language datasets, the multi-language LUKE performs quite well despite not being tailored to the legal field or any specific language. This is likely because LUKE is an innovative model specialized in the NER task. For instance, when examining the Spanish dataset, it almost matches the performance of the language-specific RoBERTa. One reason for this might be the limited number of labels in the dataset.

Table VII showcases the results obtained from training on all datasets using the multilingual LUKE.

In this table we can see how the model shows weak performance in the English dataset. This is because the English dataset has the most labels, and the majority of them are closely tied to the legal domain – an area the model wasn't trained for. During the multilingual adaptation, many of the specific legal labels were merged into more general ones, contributing to the enhanced performance. Moreover, although the mixed dataset presents same number of labels as the Spanish one, multilingual LUKE achieve better performance on the monolingual dataset.

Dataset	Model	Dev Set			
		F1 Strict	F1 Partial	F1 Exact	F1 Type Match
elenanereiss/german-ler	elenanereiss/bert-german-ler	95.77%	97.19%	96.12%	97.63%
	studio-ousia/mluke-base	83.58%	87.26%	83.94%	89.90%
Legal-ES	PlanTL-GOB-ES/roberta-base-bne-capitel-ner	92.02%	94.28%	92.30%	95.53%
	studio-ousia/mluke-base	90.50%	93.50%	90.84%	94.98%

TABLE VI: Results of L-NER task with the monolingual datasets with both language specific and multilingual models.

Dataset	Dev Set			
	F1 Strict	F1 Partial	F1 Exact	F1 Type Match
multilingual dataset	79.03%	85.18%	79.62%	88.62%
english dataset	72.49%	81.59%	74.48%	83.82%
elenanereiss/german-ler	83.58%	87.26%	83.94%	89.90%
Legal-ES	90.50%	93.50%	90.84%	94.98%

TABLE VII: Results of L-NER task using multilingual LUKE.

Model	Dev Set			
	F1 Strict	F1 Partial	F1 Exact	F1 Type Match
BERT large (ft on NER) [1]	83.01%	88.58%	83.98%	90.25%
LegalBERT-base [4]	86.84%	91.47%	87.80%	93.18%
BERT-base (ft on ECHR) [3]	85.70%	91.23%	87.27%	92.17%
LUKE-large [6]	**87.97%**	**92.25%**	**88.83%**	**93.60%**

TABLE VIII: Results of L-NER task on the English dataset.

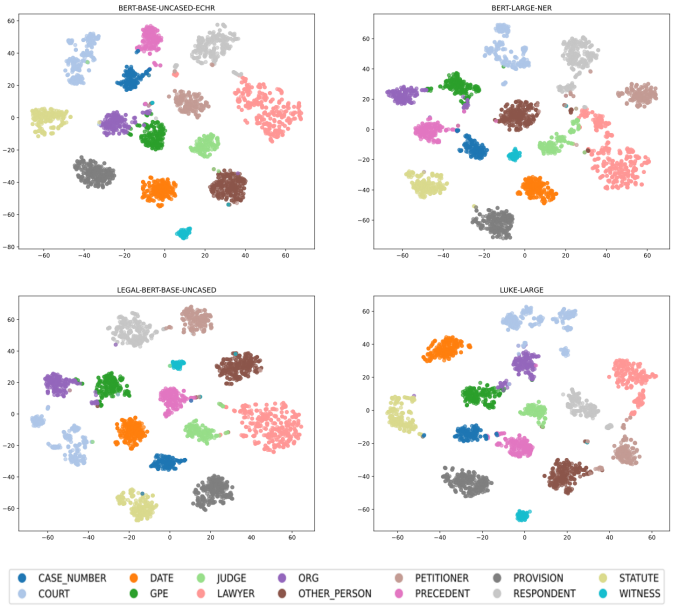

Fig. 1: Two-dimensional t-SNE visualizations. The embeddings are obtained from BERT large (ft on NER), LegalBERT-base, BERT-base (ft on ECHR) and LUKE-large, respectively. Note that *ft* indicates the fine-tuning step.

The t-SNE visualization is employed to project the entity representations into a two-dimensional space. The embeddings examined are those obtained from four different models trained on the English language dataset VIII. As already deduced in [8], LUKE-large achieve the best performance on the development set. Focusing on the visualization of the entity representations obtained 1, we can observe how each model presents significant differentiation in entity distributions across categories and aggregation of similar entity types. In particular, the LUKE-large representation exhibits almost no overlap between entity clusters. Instead the small overlaps observed appear to involve entities corresponding to more general labels (such as PER): for instance, the blurred distinction between judge, lawyer, and respondent for BERT-LARGE or between Organization and GPE for Legal BERT.

IV. Conclusion

Our study proposes a practical approach for the creation of a multilingual dateset suitable for addressing Legal Named Entity Recognition (L-NER) tasks, utilizing multilingual models and, additionally provides a detailed analysis of the performances in different scenarios.

The multilingual LUKE model proved to have robust performance on diverse language-specific datasets without being specifically tailored to the legal domain. Notably, it almost matched the Spanish language-specific model, highlighting its versatility in handling generic labels in different languages.

After training the Multilingual LUKE on the newly created multilingual Legal NER dataset we contributed to the creation of a more versatile model capable of navigating legal contexts across languages.

The English dataset's richness in legal labels challenged the multilingual LUKE model, emphasizing the need for optimization when dealing with datasets with numerous and domain-specific labels.

Moreover, previous analyses highlighted how visualization techniques which includes dimensionality reduction, as the t-SNE, can provide useful insights about models output, improving the model interpretability and revealing possible patterns in the categorization.

A major challenge in the project development was finding suitable datasets with labels that could, if properly aligned, express comparable concepts. A future challenge could be the creation of a multilingual dataset which includes a broader range of languages, while enabling a fine-grained label alignment.

REFERENCES

[1] dslim/bert-large-ner transformer model. https://huggingface.co/dslim/bert-large-NER. Accessed: 2024-01-25.

[2] elenanereiss/bert-german-ler transformer model. https://huggingface.co/elenanereiss/bert-german-ler. Accessed: 2024-01-25.

[3] nlpaueb/bert-base-uncased-echr transformer model. https://huggingface.co/nlpaueb/bert-base-uncased-echr. Accessed: 2024-01-25.

[4] nlpaueb/legal-bert-base-uncased transformer model. https://huggingface.co/nlpaueb/legal-bert-base-uncased. Accessed: 2024-01-25.

[5] Plantl-gob-es/roberta-base-bne-capitel-ner transformer model. https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-ner. Accessed: 2024-01-25.

[6] studio-ousia/luke-large transformer model. https://huggingface.co/studio-ousia/luke-large. Accessed: 2024-01-25.

[7] studio-ousia/mluke-base transformer model. https://huggingface.co/studio-ousia/mluke-base. Accessed: 2024-01-25.

[8] Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Elena Baralis, Luca Cagliero, and Francesco Tarasconi. PoliToHFI at SemEval-2023 task 6: Leveraging entity-aware and hierarchical transformers for legal entity recognition and court judgment prediction. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1401–1411, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Sau Meng Karn, and Vivek Raghavan. Named entity recognition in indian court judgments. *ArXiv*, abs/2211.03442, 2022.

[10] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. A dataset of German legal documents for named entity recognition. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France, May 2020. European Language Resources Association.

[11] Doaa Samy. Reconocimiento y clasificación de entidades nombradas en textos legales en español. In *Procesamiento del Lenguaje Natural - Nº 67*, pages 103–114. Sociedad Española para el Procesamiento del Lenguaje Natural, 2021.

[12] Doaa Samy, Jerónimo Arenas-García, and David Pérez-Fernández. Legal-ES: A set of large scale resources for Spanish legal text processing. In Doaa Samy, David Pérez-Fernández, and Jerónimo Arenas-García, editors, *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 32–36, Marseille, France, May 2020. European Language Resources Association.

[13] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Conference on Empirical Methods in Natural Language Processing*, 2020.