Deep learning techniques for Named Entity Recognition applied to Brazilian legal documents

Luan Coelho Vieira da Silva[1,*]

[1]COPPE – Graduate School and Research in Engineering, Centro de Tecnologia - Rua Horácio Macedo, Bloco G, 2030 - 101, Cidade Universitária da Universidade Federal do Rio de Janeiro, Rio de Janeiro - RJ, 21941-450, Brazil
[*]luan@cos.ufrj.br

Abstract

This article explores the application of deep learning techniques for named entity recognition (NER) in Brazilian legal documents, which is an essential task in Natural Language Processing (NLP). This study considers two domain-specific entities: legal cases and legislation. To evaluate the feasibility of applying NER to Brazilian legal texts, we employ three deep learning approaches (BiLSTM, transition-based parser and BERT). We use a data set called LeNER-Br, which consists of manually annotated legal documents in Brazilian Portuguese. Since domain-specific datasets in Brazilian Portuguese are not widely available, this dataset is particularly valuable. The performance of the models in the test split of the LeNER-Br dataset is evaluated using precision, recall, and F1-score metrics. The outcomes underscore the efficacy of deep learning models in accurately identifying both conventional and domain-specific named entities within Brazilian legal documents. Specifically, the BERT model achieves a macro average F1-score of 0.84, while the spaCy model achieves 0.83.

1 Introduction

Named Entity Recognition (NER) (Nadeau and Sekine, 2007) is a fundamental task in Natural Language Processing (NLP) that involves identifying and classifying mentions of specific types of entity, such as persons, locations, organizations, and more, within a given text. NER serves as more than just an objective in itself, it also plays a pivotal role in numerous NLP applications such as information retrieval and question answering.

The process of NER revolves around locating, extracting, and categorizing named entities in natural language texts. These named entities are objects that can be uniquely identified by proper nouns and belong to predefined classes, such as persons, locations, organizations, and even temporal and numeric expressions like dates and monetary values.

Deep Learning (DL) based models (Li et al., 2020) are able to learn complex features via non-linear activation functions. State-of-the-art DL models have the ability to capture contextual information, which is crucial for NER, as surrounding words and phrases often influence the correct identification and classification of named entities.

In this article, we leveraged three DL models for the task of recognizing named entities in Portuguese Brazilian legal documents. We chose a carefully curated dataset consisting entirely of manually annotated legal documents in Brazilian Portuguese. By focusing specifically on legal texts, we aim to verify the feasibility of applying NER to lawful texts from Brazil considering the unique challenges that arise in this domain due to textual peculiarities (Luz de Araujo et al., 2018).

2 Related works

Significant progress has been achieved in NER applied to legal texts. Dozier et al. (2010) proposes tags for legal documents such as US case law, depositions, and pleadings, with entities including document type, jurisdiction, court, and judge. Similar works with proposed frameworks have been applied in other languages, such as Glaser, Waltl, and Matthes (2018). Portuguese named entity recognition datasets are scarce, HAREM (Freitas et al., 2010) and WikiNER (Nothman et al., 2013) are regarded as two of the most relevant ones.

At the time of writing their article, the authors of the LeNER-Br (Luz de Araujo et al., 2018) were not aware of any other dataset in Portuguese consisting of legal documents nor a baseline method for NER applied in such documents. Furthermore, our search did not yield any prior nor subsequent Portuguese dataset in this domain. In the same work they also propose a LSTM-CRF model and evaluate their results. Nonetheless, after their publication, others have built on their work. Wang et al. (2020) also put forward frameworks for recognizing entities in legal documents written in Portuguese using a methodology similar to that employed in the present article.

3 LeNER-Br Dataset

LeNER-Br (Luz de Araujo et al., 2018) is a manually annotated Portuguese language data set for the recognition of named entities, specifically focused on legal documents. It includes named entities related to legislations and legal cases, aiming to enhance the extraction of lawful knowledge.

To construct the dataset, 66 legal documents were gathered from various

Brazilian courts, covering both the superior and state levels. This collection included documents from institutions such as the *Supremo Tribunal Federal* (Supreme Federal Court), *Superior Tribunal de Justiça* (superior Court of Justice), *Tribunal de Justiça de Minas Gerais* (Justice Court of Minas Gerais State), and *Tribunal de Contas da União* (Federal Court of Accounts). Additionally, four other legislation documents were included, resulting in a total of 70 documents. For each document, the NLTK library (Bird, Klein, and Loper, 2009) was used to split the text into a list of sentences and tokenize them.

Table 1 shows the number of documents, sentences, and tokens in each data split. Table 2 presents the number of tokens for each of the entities in the data set. Entities are assumed to belong to a single sentence. Most tokens do not have an associated entity.

Table 1: Set configuration (adapted from Luz de Araujo et al. (2018))

Set	Documents	Sentences	Tokens
Training	50	7,827	229,277
Validation	10	1,176	41,166
Test	10	1,389	47,630

Table 2: Distribution of Entities per Token and Data Split (adapted from Luz de Araujo et al. (2018))

Entity	Training	Validation	Test
Legal Cases	3,967	743	660
Legislation	13,039	2,609	2,669
Location	1,417	244	132
Organization	6,671	1,608	1,367
Person	4,612	894	735
Time	2,343	543	260
No entity	197,228	34,525	41,807

4 Bidirectional LSTM-RNN

A bidirectional LSTM-RNN (BiLSTM) architecture (Lyu et al., 2017; Sherstinsky, 2020) was implemented using the Keras (Chollet et al., 2015) library, which is a high-level deep learning framework built on top of TensorFlow (Abadi et al., 2016). The model consists of an input layer, an embedding layer to represent the tokens, a bidirectional LSTM layer to capture contextual information, and a time-distributed dense layer for sequence labeling.

Using a bidirectional approach, the model is able to capture information from both past and future contexts, enhancing its understanding of sequential data. This bidirectional nature enables the model to make more informed predictions by considering the complete context surrounding each token in the sequence.

The resulting model has a total of 2,854,925 parameters. This large number of parameters indicates the model's capacity to learn complex patterns and relationships within the data and is characteristic of deep learning models. However, it is important to note that a higher number of parameters also increases the risk of overfitting when training with limited data. Table 3 presents a list of hyperparameters of the model that were tuned to optimize the performance of the model on the task.

Table 3: Bidirectional LSTM-RNN hyperparameters

Hyperparameter	**Value**
Embedding Dimension	128
LSTM Units	64
Number of Epochs	10
Optimizer	Adam
Loss Function	Categorical Cross-Entropy

5 spaCy

spaCy (Honnibal et al., 2020) is a free and open-source library designed specifically for advanced Natural Language Processing (NLP) tasks in Python, capable of processing large volumes of text. The model architecture follows a transition-based parser approach (Explosion, 2023a), with a tok2vec module for tokenization and word embeddings. The tok2vec module utilizes a Tok2VecListener sublayer, which receives the output from a MaxoutWindowEncoder. Tok2vec stands for "token to vector" and involves transforming tokens into continuous vector representations using neural networks.

spaCy first segments text into tokens to create a library-specific Doc variable (Explosion, 2023b). This Doc is then processed in several different steps in the processing pipeline, separated by components. At the end of each step, a new Doc file is returned and fed to the next component, resulting in a final Doc at the end of the pipeline. Figure 1 shows an example of this iteration, in which the tagger component assigns part-of-speech tags, the parser creates dependency labels, and the NER detects and labels named entities.

The spaCy training process involves an iterative approach in which the model predictions are compared with reference annotations to estimate the loss gradient. This gradient is then utilized in the backpropagation to calculate the weight gradients. By updating the weights based on these gradients, the model's predictions gradually align more closely with the reference labels over time. The

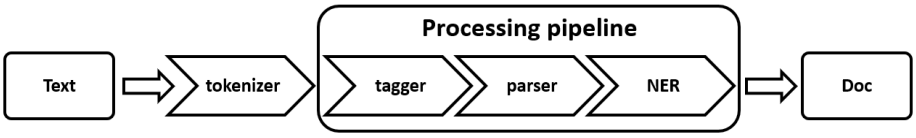

Figure 1: Spacy Pipeline, adapted from Explosion, 2023b

documentation (Explosion, 2023c) offers training instructions that can be optimized for accuracy and performance.

We used pt_core_news_lg (Explosion, 2023d), a Portuguese pipeline trained in the Named Entity Recognition task. This pipeline contains 500,000 unique vectors. Subsequently, we utilized our domain-specific training and validation datasets to construct a customized model incorporating our predefined entities. The model hyperparameters are shown in Table 4.

Table 4: SPACY hyperparameters

Hyperparameter	**Value**
Hidden Width	64
Maxout Pieces	2
Learning Rate	0.001
L2 Weight Decay	0.01
Dropout Rate	0.1
Batcher Start Size	100
Batcher Stop Size	1000
Batcher Compound Factor	1.001
Maximum Training Steps	20000
Evaluation Frequency	200

6 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Devlin et al., 2018) is a deep learning model designed to pretrain bidirectional representations from unlabeled text, considering both left and right context across all layers. This pretrained BERT model can be fine-tuned with an additional output layer to achieve state-of-the-art performance on various tasks, including named entity recognition (NER).

In this study, we employed BERTimbau (Souza, Nogueira, and Lotufo, 2019; Souza, Nogueira, and Lotufo, 2020), a pretrained BERT model specifically tailored for Brazilian Portuguese. BERTimbau has demonstrated exceptional results in NER tasks, making it an ideal choice for our research. We fine-tuned

the BERT-Base cased architecture, which consists of 12 layers and 110 million parameters, using the LeNER-Br dataset. The model's hyperparameters are presented in Table 5.

Table 5: BERT hyperparameters

Hyperparameter	**Value**
Learning Rate	0.00002
Training Batch Size	4
Evaluation Batch Size	4
Optimizer	Adam
Scheduler Type	linear
Number of Epochs	15

7 Evaluation Metrics

In the context of multilabel classification for Named Entity Recognition (NER) tasks, for a given label, we define the following terms:

- True Positive (TP): Instances where the model correctly predicts the named entity label.
- False Positive (FP): Instances where the model incorrectly predicts the named entity label when it is not the actual label.
- False Negative (FN): Instances where the model incorrectly predicts the absence the label when it is the actual label.

These terms allow us to calculate our evaluation metrics, which are commonly used to evaluate the performance of models in token classification tasks. Recall measures the ability of the model to correctly identify a named entity label out of all the actual positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision measures the ability of the model to correctly identify a named entity label out of all the instances predicted as positive. It is calculated as the ratio of true positives to the sum of true positives and false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of both metrics.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

8 Results

We evaluated all of the proposed models using the test split of the LeNER-Br data set, considering the metrics discussed earlier. Most tokens in the data set have no associated entity, as indicated in Table 2. Notably, Time and Location entities have fewer tokens associated with them compared to other entity types. Consequently, there is a higher risk of misclassification, since the model must accurately identify the relatively scarce tokens that contain these specific entities. Detailed results can be found in Table 6.

Table 6: Entity Recognition Evaluation

	BiLSTM			**spaCy**			**BERT**			
Entity	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
Legal cases	0.78	0.57	0.66	0.86	0.67	0.75	**0.87**	**0.88**	**0.87**	660
Legislation	0.94	0.88	0.91	**0.95**	**0.97**	**0.96**	0.94	0.95	0.94	2669
Location	0.56	0.34	0.42	**0.57**	0.64	0.60	0.53	**0.73**	**0.62**	132
Organization	0.86	0.75	0.80	**0.88**	0.75	**0.81**	0.80	**0.82**	**0.81**	1367
Person	**0.95**	0.72	0.82	0.93	**0.96**	0.94	0.94	**0.96**	**0.95**	735
Time	0.83	0.49	0.62	**0.88**	**0.90**	**0.89**	0.84	0.88	0.86	260
Macro Average	0.82	0.62	0.70	**0.85**	0.82	0.83	0.82	**0.87**	**0.84**	5823

During our experimentation, it became evident that our BiLSTM model exhibited inferior performance compared to the other models, particularly in terms of the recall metric, indicating a higher number of false negatives. A study conducted by Patel, Nguyen, and Baraniuk (2016) delves into the concept of generalization error that can occur in nonprobabilistic deep learning models, highlighting how multiple runs of such models can yield varying outcomes. Nevertheless, it is important to acknowledge that even in this model, the recognition of legislative and legal cases entities aligned with the results obtained for other entity types, which are commonly addressed in named entity recognition tasks and extensively discussed in the literature.

Furthermore, legislation demonstrated the highest performance, as anticipated considering its larger number of observations, but still serving as a reliable indicator of the applicability of NER to legal documents incorporating custom entities. In terms of precision, spaCy outperformed other models, which indicates that it had fewer false positives and more accurately identified the relevant entities, while BERT exhibited better recall performance, which means it had fewer false negatives and was more effective at capturing all the entities.

9 Conclusion

This study used three different models: a bidirectional LSTM-RNN (BiLSTM), a transition-based parser using the spaCy library, and the Bidirectional Encoder Representations from Transformers (BERT) model. These results demonstrate the effectiveness of deep learning models in capturing contextual information and identifying named entities regarding legislation and legal cases in Brazilian legal documents. The LeNER-Br dataset, consisting of manually annotated legal documents in Brazilian Portuguese, provided a valuable resource for training and evaluating the models.

A significant obstacle in NER involves the scarcity of manually annotated datasets, particularly domain-specific and in Brazilian Portuguese, as in the present study. Automatically labeled corpora suffer from inferior annotation quality. Future endeavors should focus on integrating additional datasets that cover a wider range of legislation. Additionally, investigating the impact of different hyperparameter values on our models and incorporating a probabilistic framework capable of capturing classification probabilities would be crucial areas for further exploration.

References

Abadi, Martín et al. (2016). "Tensorflow: a system for large-scale machine learning." In: *Osdi*. Vol. 16. 2016. Savannah, GA, USA, pp. 265–283.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Chollet, François et al. (2015). *Keras*. `https://keras.io`.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dozier, Christopher et al. (2010). *Named entity recognition and resolution in legal text*. Springer.

Explosion (2023a). *Parsing English in Python*. `https://explosion.ai/blog/parsing-english-in-python`. Accessed on June 23, 2023.

— (2023b). *spaCy - Processing Pipelines*. `https://spacy.io/usage/processing-pipelines`. Accessed on June 23, 2023.

— (2023c). *spaCy - Training a Named Entity Recognizer*. `https://spacy.io/usage/training`. Accessed on June 23, 2023.

— (2023d). *spacy-models: Release pt_core_news_lg-3.5.0*. `https://github.com/explosion/spacy-models/releases/tag/pt_core_news_lg-3.5.0`. Accessed on June 23, 2023.

Freitas, Cláudia et al. (2010). "Second HAREM: advancing the state of the art of named entity recognition in Portuguese". In: *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta*

17-23 May de 2010) European Language Resources Association. European Language Resources Association.

Glaser, Ingo, Bernhard Waltl, and Florian Matthes (2018). "Named entity recognition, extraction, and linking in german legal contracts". In: *IRIS: Internationales Rechtsinformatik Symposium*, pp. 325–334.

Honnibal, Matthew et al. (2020). "spaCy: Industrial-strength Natural Language Processing in Python". In: DOI: 10.5281/zenodo.1212303.

Li, Jing et al. (2020). "A survey on deep learning for named entity recognition". In: *IEEE Transactions on Knowledge and Data Engineering* 34.1, pp. 50–70.

Luz de Araujo, Pedro Henrique et al. (2018). "LeNER-Br: a dataset for named entity recognition in Brazilian legal text". In: *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*. Springer, pp. 313–323.

Lyu, Chen et al. (2017). "Long short-term memory RNN for biomedical named entity recognition". In: *BMC bioinformatics* 18, pp. 1–11.

Nadeau, David and Satoshi Sekine (2007). "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Nothman, Joel et al. (2013). "Learning multilingual named entity recognition from Wikipedia". In: *Artificial Intelligence* 194, pp. 151–175.

Patel, Ankit B, Minh T Nguyen, and Richard Baraniuk (2016). "A probabilistic framework for deep learning". In: *Advances in neural information processing systems* 29.

Sherstinsky, Alex (2020). "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404, p. 132306.

Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo (2019). "Portuguese Named Entity Recognition using BERT-CRF". In: *arXiv preprint arXiv:1909.10649*. URL: http://arxiv.org/abs/1909.10649.

— (2020). "BERTimbau: Pretrained BERT Models for Brazilian Portuguese". In: *Intelligent Systems*. Ed. by Ricardo Cerri and Ronaldo C. Prati. Cham: Springer International Publishing, pp. 403–417. ISBN: 978-3-030-61377-8.

Wang, Zhili et al. (2020). "Named entity recognition method of brazilian legal text based on pre-training model". In: *Journal of Physics: Conference Series*. Vol. 1550. 3. IOP Publishing, p. 032149.

Supporting Documents

All resources used in this article, including the Python codes and link to the Overleaf .tex file are available in the Google Drive folder that can be accessed at https://drive.google.com/drive/folders/1kcIvQbaa1OjMplxGPFM0QA58vlR5XgDI?usp=sharing.