LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text

Pedro Henrique Luz de Araujo[1], Teófilo E. de Campos[2],
Renato R. R. de Oliveira[1], Matheus Stauffer[1], Samuel Couto[1], and
Paulo Bermejo[1]

[1] NEXT, University of Brasília, Brasília, Brazil
[2] Department of Computer Science, University of Brasília, DF, Brazil
pedrohluzaraujo@gmail.com, t.decampos@st-annes.oxon.org
{renatooliveiraz, matheusstauffer, samuelcouto, paulobermejo}@next.unb.br

Abstract. Named entity recognition systems have the untapped potential to extract information from legal documents, which can improve information retrieval and decision-making processes. In this paper, a dataset for named entity recognition in Brazilian legal documents is presented. Unlike other Portuguese language datasets, this dataset is composed entirely of legal documents. In addition to tags for persons, locations, time entities and organizations, the dataset contains specific tags for law and legal cases entities. To establish a set of baseline results, we first performed experiments on another Portuguese dataset: Paramopama. This evaluation demonstrate that LSTM-CRF gives results that are significantly better than those previously reported. We then retrained LSTM-CRF, on our dataset and obtained F_1 scores of 97.04% and 88.82% for Legislation and Legal case entities, respectively. These results show the viability of the proposed dataset for legal applications.

Keywords: Named Entity Recognition · Natural Language Processing · Portuguese Processing

1 Introduction

Named entity recognition (NER) is the process of finding, extracting and classifying named entities in natural language texts. Named entities are objects that can be designated by a proper noun and fit predefined classes such as persons, locations and organizations. In addition, the NER community has found useful to include temporal and numeric expressions (e.g. dates and monetary values) as named entities [18].

The state-of-the-art entity recognition systems [13, 14] are based on Machine Learning techniques, employing statistical models that need to be trained on a large amount of labeled data to achieve good performance and generalization capabilities [15]. The process of labeling data is expensive and time consuming since the best corpora are manually tagged by humans.

There are few manually annotated corpora in Portuguese. Some examples are the first and second HAREM [22, 5] and Paramopama [17]. Another approach is to automatically tag a corpus, like the one proposed in [19] that originated the WikiNER corpus. Such datasets have lower quality than manually tagged ones, as they do not take into consideration sentence context, which can result in inconsistencies between named entity categories [17].

An area that can potentially leverage the information extraction capabilities of NER is the judiciary. The identification and classification of named entities in legal texts, with the inclusion of juridical categories, enable applications such as providing links to cited laws and legal cases and clustering of similar documents.

There are some issues that discourage the use of models trained on existing Portuguese corpora for legal text processing. Foremost, legal documents have some idiosyncrasies regarding capitalization, punctuation and structure. This particularity can be exemplified by the excerpts below:

> EMENTA: APELAÇÃO CÍVEL - AÇÃO DE INDENIZAÇÃO POR DANOS MORAIS - PRELIMINAR - ARGUIDA PELO MINISTÉRIO PÚBLICO EM GRAU RECURSAL - NULIDADE - AUSÊNCIA DE INTERVENÇÃO DO PARQUET NA INSTÂNCIA A QUO - PRESENÇA DE INCAPAZ - PREJUÍZO EXISTENTE - PRELIMINAR ACOLHIDA - NULIDADE RECONHECIDA.

> HABEAS CORPUS 110.260 SÃO PAULO RELATOR : MIN. LUIZ FUX PACTE.(S) :LAERCIO BRAZ PEREIRA SALES IMPTE.(S) :DEFENSORIA PÚBLICA DA UNIÃO PROC.(A/S)(ES) :DEFENSOR PÚBLICO-GERAL FEDERAL COATOR(A/S)(ES) :SUPERIOR TRIBUNAL DE JUSTIÇA

In these passages, not only are all letters capitalized, but also there is no ordinary phrase structure of subject and predicate. Intuitively, it follows that the distribution of such documents differs from the existing corpora in a way that models trained on them will perform poorly when processing legal documents. Also, as they do not have specific tags for juridical entities, the models would fail to extract such legal knowledge.

The present paper proposes a Portuguese language dataset for named entity recognition composed entirely of manually annotated legal documents. Furthermore two new categories (LEGISLACAO, for named entities referring to laws; and JURISPRUDENCIA, for named entities referring to legal cases) are added to better extract legal knowledge.

Some efforts have been made on NER in legal texts. For instance, Dozier et al. [4] propose a NER system for Title, Document Type, Jurisdiction, Court and Judge tagging. Nevertheless, only the first entity is identified using a statistical approach, while the others are classified with contextual rules and lookup tables. Cardellino et al. [2] used the Wikipedia to generate an automatically annotated corpus, tagging persons, organizations, documents, abstraction (rights, legal doctrine) and act (statutes) entities. As far as we are aware, the present paper is the first to propose a benchmark dataset and a baseline method for NER in legal texts in Portuguese.

Table 1. Sentence, token and document count for each set.

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

The rest of this paper is organized as follows. In section 2, we discuss the dataset creation process. We present the model used to evaluate our dataset in section 3, along with the training of the model and our choice of hyperparameters in section 4. In section 5 we present the results achieved regarding the test sets and section 6 presents the final considerations.

2 The LeNER-Br dataset

To compose the dataset, 66 legal documents from several Brazilian Courts were collected. Courts of superior and state levels were considered, such as Supremo Tribunal Federal, Superior Tribunal de Justiça, Tribunal de Justiça de Minas Gerais and Tribunal de Contas da União. In addition, four legislation documents were collected, such as Lei Maria da Penha, giving a total of 70 documents.

For each document, the NLTK [1] was used to split the text into a list of sentences and tokenize them. The final output for each document is a file with one word per line and an empty line delimiting the end of a sentence.

After preprocessing the documents, WebAnno [3] was employed to manually annotate each one of the documents with the following tags: "ORGANIZACAO" for organizations, "PESSOA" for persons, "TEMPO" for time entities, "LOCAL" for locations, "LEGISLACAO" for laws and "JURISPRUDENCIA" for decisions regarding legal cases. The last two refer to entities that correspond to "Act of Law" and "Decision" classes from the Legal Knowledge Interchange Format ontology [10] respectively.

The IOB tagging scheme [21] was used, where "B-" indicates that a tag is the beginning of a named entity, "I-" indicates that a tag is inside a named entity and "O-" indicates that a token does not pertain to any named entity. Named entities are assumed to be non-overlapping and not spanning more than one sentence.

To create the dataset, 50 documents were randomly sampled for the training set and 10 documents for each of the development and test sets. The total number of tokens in LeNER-Br is comparable to other named entity recognition corpora such as Paramopama and CONLL-2003 English [24] datasets (318,073, 310,000 and 301,418 tokens respectively). Table 1 presents the number of tokens and sentences of each set and Table 2 displays the number of words in named entities of each set per class. Table 3 presents an excerpt from the training set.

Table 2. Named entity word count for each set.

Category	Training set	Development set	Test set
Person	4,612	894	735
Legal cases	3,967	743	660
Time	2,343	543	260
Location	1,417	244	132
Legislation	13,039	2,609	2,669
Organization	6,671	1,608	1,367

3 The baseline model: LSTM-CRF

To establish a methodological baseline on our dataset, we chose the LSTM-CRF model, proposed in [13]. This model is proven to be capable of achieving state-of-the-art performance on the English CoNLL-2003 test set [24] (a F1-score of 90.94%). It also has readily available open source implementations [6], which was adapted for the needs of the present work.

The architecture of the model consists of a Bidirectional [7] Long Short-Term Memory Layer (LSTM) [9] followed by a CRF [12] layer. The input of the model is a sequence of vector representations of individual words constructed from the concatenation of both word embeddings and character level embeddings.

For the word lookup table we used 300 dimensional GloVe [20] word embeddings pretrained on a multi-genre corpus formed by both Brazilian and European Portuguese texts [8]. These word embeddings are fine tuned during training.

The character level embeddings are obtained from a character lookup table initialized at random values with embeddings for every character in the dataset. The embeddings are fed to a separate bidirectional LSTM layer. The output is then concatenated with the pretrained word embeddings, resulting in the final vector representation of the word. Figure 1 presents an overview of this process.

To reduce overfitting and improve the generalization capabilities of the model a dropout mask [23] is applied to the outputs of both bidirectional LSTM layers, i.e., the one following the character embeddings and the one after the final word representation. Figure 2 shows the main architecture of the model.

4 Experiments and hyper-parameters setting

This section presents the methods employed to train the model and displays the hyper-parameters that achieved the best performance.

Both Adam [11] and Stochastic Gradient Descent (SGD) with momentum were evaluated as optimizers. Although SGD had slower convergence, it achieved better scores than Adam. Gradient clipping was employed to prevent the gradients from exploding.

After experimenting with hyper-parameters, the best performance was achieved with the ones used in [13], presented in table 4. It is worth noting that the number of LSTM units refers to one direction only. Since the LSTM are bidirectional,

Table 3. Two excerpts from the training set. Each line has a word, a space delimiter and the tag corresponding to the word. Sentences are separated by an empty line.

A	O	TJMG	B-ORGANIZACAO
falta	O	-	O
de	O	Apelação	B-JURISPRUDENCIA
intervenção	O	Cível	I-JURISPRUDENCIA
do	O	1.0549.15.003028-2/003	I-JURISPRUDENCIA
Ministério	B-ORGANIZACAO	,	O
Público	I-ORGANIZACAO	Relator	O
nas	O	(	O
ações	O	a	O
em	O	)	O
que	O	:	O
deva	O	Des	O
figurar	O	.	O
como	O	(	O
fiscal	O	a	O
da	O	)	O
lei	O	Otávio	B-PESSOA
e	O	Portes	I-PESSOA
da	O	,	O
Constituição	B-LEGISLACAO	16ª	B-ORGANIZACAO
(	O	CÂMARA	I-ORGANIZACAO
custus	O	CÍVEL	I-ORGANIZACAO
legis	O	,	O
et	O	julgamento	O
constituitionis	O	em	O
’	O	28/09/2017	B-TEMPO
)	O	,	O
enseja	O	publicação	O
de	O	da	O
forma	O	súmula	O
inexorável	O	em	O
a	O	06/10/2017	B-TEMPO
nulidade	O	)	O
do	O	Assim	O
processo	O	sendo	O
,	O	,	O
segundo	O	entendo	O
prescreve	O	que	O
o	O	deve	O
artigo	B-LEGISLACAO	ser	O
279	I-LEGISLACAO	acolhida	O
...	...	...	...

the final number of units doubles. Moreover, the learning rate decay is applied after every epoch. The net parameters were saved only when achieving better performance on the validation set than past epochs.

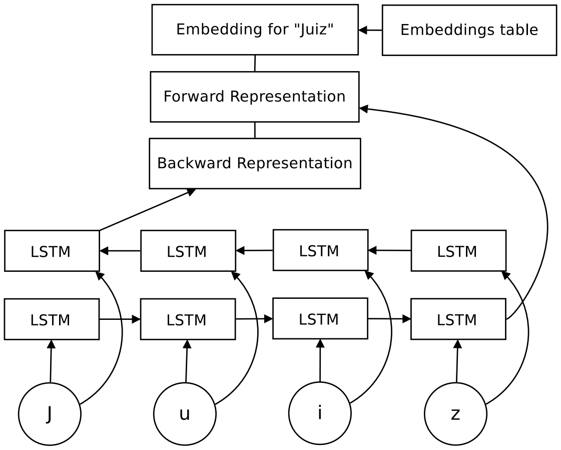

Fig. 1. Each word vector representation is a result of the concatenation of the outputs of a bidirectional LSTM and the word level representation from the word lookup table.

Table 4. Model hyper-parameter values.

Hyper-parameter	Value
Word embedding dimension	300
Character embedding dimension	50
Number of epochs	55
Dropout rate	0.5
Batch size	10
Optimizer	SGD
Learning rate	0.015
Learning rate decay	0.95
Gradient clipping threshold	5
First LSTM layer hidden units	25
Second LSTM layer hidden units	100

The model was first trained using the Paramopama Corpus [17] to evaluate if it could achieve state-of-the-art performance on a Portuguese dataset. This dataset contains four different named entities: persons, organizations, locations and time entities. After confirming that the model performed better than the state-of-the-art model (ParamopamaWNN [16]), the LSTM-CRF network was trained with the proposed dataset.

The preprocessing steps applied were lowercasing the words and replacing every digit with a zero. Both steps are necessary to match the preprocessing of the pretrained word embeddings. Since the character-level representation preserves the capitalization, this information is not lost when the words are lowercased.

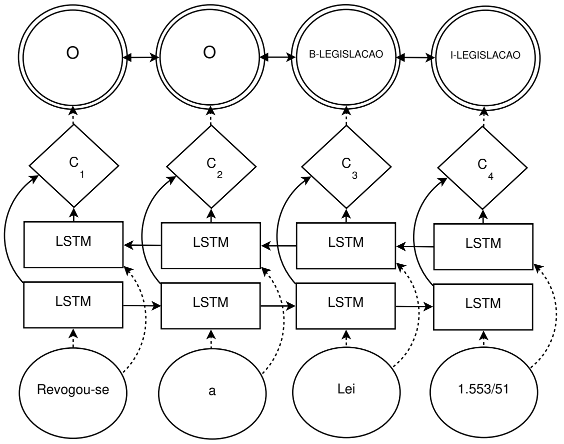

Fig. 2. The LSTM-CRF model. The word vector representations serve as input to a bidirectional LSTM layer. C_i represents the concatenation of left and right context of word *i*. Dotted lines represent connections after a dropout layer is applied.

5 Results

The metric used to evaluate the performance of the model on both datasets was the F_1 Score. Tables 5 and 6 compare the performance of the LSTM-CRF [13] and ParamopamaWNN [16] models on different test sets. Test Set 1 and Test Set 2 are the last 10% of the WikiNER [19] and HAREM [22] corpora respectively. Table 7 shows the scores achieved by the LSTM-CRF model when training on the proposed dataset.

Table 5. Results on Paramopama Test Set 1 (10% of the WikiNER [19]).

	ParamopamaWNN			LSTM-CRF		
Entity	Precision	Recall	F_1	Precision	Recall	F_1
Person	83.76%	90.50%	87.00%	**91.80%**	**92.43%**	**92.11%**
Location	87.55%	**88.09%**	87.82%	**92.80%**	87.39%	**90.02%**
Organization	69.55%	82.35%	75.41%	**72.27%**	**83.94%**	**77.67%**
Time	86.96%	89.06%	88.00%	**92.54%**	**96.66%**	**94.56%**
Overall	86.45%	89.77%	88.08%	**90.01%**	**91.16%**	**90.50%**

The obtained results show that the LSTM-CRF network outperforms the ParamopamaWNN on both test sets, achieving better precision, recall and F_1 scores in the majority of the entities. Furthermore, it improved the overall score by 2.48% and 4.58% on the first and second test sets respectively.

As far as we are aware, there is no published material about legal entities recognition in Portuguese, so it was not possible to establish a baseline for comparison on LeNER-Br. Despite that, the obtained results on LeNER-Br show

Table 6. Results on Paramopama Test Set 2 (HAREM [22]).

	ParamopamaWNN			LSTM-CRF		
Entity	Precision	Recall	F_1	Precision	Recall	F_1
Person	84.36%	88.67%	86.46%	**94.10%**	**95.78%**	**94.93%**
Location	84.08%	86.85	85.44%	**90.51%**	**92.26%**	**91.38%**
Organization	81.48%	54.15%	65.06%	**83.33%**	**78.46%**	**80.82%**
Time	**98.37%**	87.40%	92.56%	91.73%	**94.01%**	**92.86%**
Overall	83.83%	88.65%	86.17%	**90.44%**	**91.10%**	**90.75%**

Table 7. Results on LeNER-Br, our dataset for NER on Legal texts from Brazil.

Entity	Precision	Recall	F_1
Person	94.44%	92.52%	93.47%
Location	61.24%	59.85%	60.54%
Organization	91.27%	85.66%	88.38%
Time	91.15%	91.15%	91.15%
Legislation	97.08%	97.00%	97.04%
Legal cases	87.39%	90.30%	88.82%
Overall	93.21%	91.91%	92.53%

that a model trained with it can achieve performance in legal cases and legislation recognition comparable to the ones seen in Paramopama entities, with F_1 scores of 88.82% and 97.04% respectively. In addition, person, time entities and organization classification scores were compatible with the ones observed in the Paramopama scenarios, obtaining scores greater than 80%.

However, location entities have a noticeably lower score than the others on LeNER-Br. This drop could be due to many different reasons. The most important one is probably the fact that words belonging to location entities are rare in LeNER-Br, representing 0.61% and 0.28% of the words pertaining to entities in the train and test sets respectively. Furthermore, location entities are easily mislabeled, as there are words that, depending on the context, may refer to a person, a location or a organization. A good example is treating the name of an avenue as the name of a person. For instance, instead of identifying "avenida José Faria da Rocha" as a location, the model classified "José Faria da Rocha" as a person.

6 Conclusion

This paper presented LeNER-Br, a Portuguese language dataset for named entity recognition applied to legal documents. As far as we are aware, this is the first dataset of its kind. LeNER-Br consists entirely of manually annotated legislation and legal cases texts and contains tags for persons, locations, time entities, organizations, legislation and legal cases. A state-of-the-art machine learning model, the LSTM-CRF, trained on this dataset was able to achieve a good performance:

average F_1 score of 92.53%. There is room for improvement, which means that this dataset will be relevant to benchmark methods that are sill to be proposed.

Future work would include the expansion of the dataset, adding legal documents from different courts and other kinds of legislation, e.g. Brazilian Constitution, State Constitutions, Civil and Criminal Codes, among others. In addition, the use of word embeddings trained on a large corpus of legislation and legal documents could potentially improve the performance of the model.

Acknowledgements

TEdC acknowledges the support of “Projeto de Pesquisa&Desenvolvimento de aprendizado de máquina (machine learning) sobre dados judiciais das repercussões gerais do Supremo Tribunal Federal - STF”.

References

1. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. (2009)
2. Cardellino, C., Teruel, M., Alonso Alemany, L., Villata, S.: A low-cost, high-coverage legal named entity recognizer, classifier and linker. In: Proceedints of the 16th International Conference on Artificial Intelligence and Law (ICAIL). London, United Kingdom (June 2017), preprint available from `https://hal.archives-ouvertes.fr/hal-01541446`
3. de Castilho, R.E., Mujdricza-Maydt, E., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A., Biemann, C.: A web-based tool for the integrated annotation of semantic and syntactic structures. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). pp. 76–84 (2016)
4. Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., Wudali, R.: Named entity recognition and resolution in legal text. In: Semantic Processing of Legal Texts, pp. 27–43. Springer (2010)
5. Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P.: Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In: Language Resources and Evaluation Conference (LREC). European Language Resources Association (2010)
6. Genthial, G.: Sequence tagging - named entity recognition with Tensorflow. GitHub repository `https://github.com/guillaumegenthial/sequence_tagging/tree/0048d604f7a4e15037875593b331e1268ad6e887` (2017)
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
8. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: Proceedings of Symposium in Information and Human Language Technology. Sociedade Brasileira de Computação, Uberlandia, MG, Brazil (October 2–5 2017), preprint available at `https://arxiv.org/abs/1708.06025`
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

10. Hoekstra, R., Breuker, J., Bello, M.D., Boer, A.: The LKIF Core ontology of basic legal concepts. In: Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (2007)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015), preprint available at `https://arxiv.org/abs/1412.6980`
12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning (ICML). ACM (2001), `http://portal.acm.org/citation.cfm?id=655813`
13. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT. pp. 260–270. Association for Computational Linguistics (ACL), San Diego, California (June 12-17 2016), preprint available at `https://arxiv.org/abs/1603.01360`
14. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 1064–1074. ACL, Berlin, Germany (August 7-12 2016), preprint available at `https://arxiv.org/abs/1603.01354`
15. Mansouri, A., Affendey, L.S., Mamat, A.: Named entity recognition approaches. International Journal of Computer Science and Network Security **8**(2), 339–344 (2008)
16. Mendonça Jr., C.A.E., Barbosa, L.A., Macedo, H.T., São Cristóvão, S.: Uma arquitetura híbrida LSTM-CNN para reconhecimento de entidades nomeadas em textos naturais em língua portuguesa. In: XIII Encontro Nacional de Intelig^encia Artificial e Computacional (ENIAC). SBC (2016)
17. Mendonça Jr., C.A.E., Macedo, H., Bispo, T., Santos, F., Silva, N., Barbosa, L.: Paramopama: a Brazilian-Portuguese corpus for named entity recognition. In: XII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). SBC (2015)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
19. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence **194**, 151–175 (2013)
20. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
21. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Natural language processing using very large corpora, pp. 157–176. Springer (1999). https://doi.org/10.1007/978-94-017-2390-9_10, preprint available at `http://arxiv.org/abs/cmp-lg/9505040`
22. Santos, D., Cardoso, N.: A golden resource for named entity recognition in Portuguese. In: International Workshop on Computational Processing of the Portuguese Language. pp. 69–79. Springer (2006)
23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
24. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL. vol. 4, pp. 142–147. Association for Computational Linguistics (2003)