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]
Paulo H. S. 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*
`http://www.cic.unb.br/~teodecampos/LeNER-Br/`

September 26, 2018

Overview

Introduction

Named Entity Recognition

- Named Entity Recognition (NER) is a sub-task of Information Extraction (IE) that aims to find, extract and classify named entities in natural language text.
- Pedro, who studies at Universidade de Brasília, went to Canela in 2018.

Named Entity Recognition

Entities

- Person
- Time
- Organization
- Location

Pedro, who studies at Universidade de Brasília, went to Canela in 2018.

NER in legal texts

- NER can help with the processing of legal text.
- By adding domain-specific entities such as "legislation" and "legal cases" we can enable application such as:
  - Linking case and law citations to the relevant documents;
  - Clustering similar documents;
- Unfortunately, some issues discourage the use of models trained on existing Portuguese corpora for legal text processing:
  - Capitalization
  - Punctuation
  - Phrase structure
  - No domain-specific entities

NER in legal texts

Excerpt 1

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.

Excerpt 2

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

LeNER-Br

- In addition to general entities (person, location, time and organization):
  - LEGISLACAO, for law citations;
  - JURISPRUDENCIA, for legal cases citations;
- Composed entirely of manually annotated legal texts.

The LeNER-Br dataset

The LeNER-Br dataset

- Composed of 70 texts:
  - 66 legal documents from several Brazilian Courts;
  - 4 legislation documents;
- NLTK was used to tokenize and split each text into a list of sentences.
- Each of the documents was then annotated with the following tags:
  - ORGANIZACAO for organizations;
  - PESSOA for persons;
  - TEMPO for time entities;
  - LOCAL for locations;
  - LEGISLACAO for laws;
  - JURISPRUDENCIA for legal decisions;
- The IOB tagging scheme was employed, where "B-" is used for tags that begin named entities, "I-" for tags inside named entities and "O" for those that don't belong to any entity.

The LeNER-Br dataset

- The documents were randomly split:
  - 50 for training set;
  - 10 for validation set;
  - 10 for test set;

Table: 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 LeNER-Br dataset

Table: Named entity word count for each set.

Category	Training set	Development set	Test set
Person	4,612	894	735
Legal decision	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

The LeNER-Br dataset

Table: 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.

...	...
lei	O
e	O
da	O
Constituição	B-LEGISLACAO
(	O
custus	O
legis	O
et	O
constituitionis	O
'	O
...	...

...	...
Otávio	B-PESSOA
Portes	I-PESSOA
,	O
16ª	B-ORGANIZACAO
CÂMARA	I-ORGANIZACAO
CÍVEL	I-ORGANIZACAO
,	O
julgamento	O
em	O
28/09/2017	B-TEMPO
...	...

The baseline model: LSTM-CRF

The baseline model: LSTM-CRF

- The architecture was shown to achieve state-of-the-art performance on German, English, Dutch and Spanish datasets [Lample, Guillaume et al, 2016].

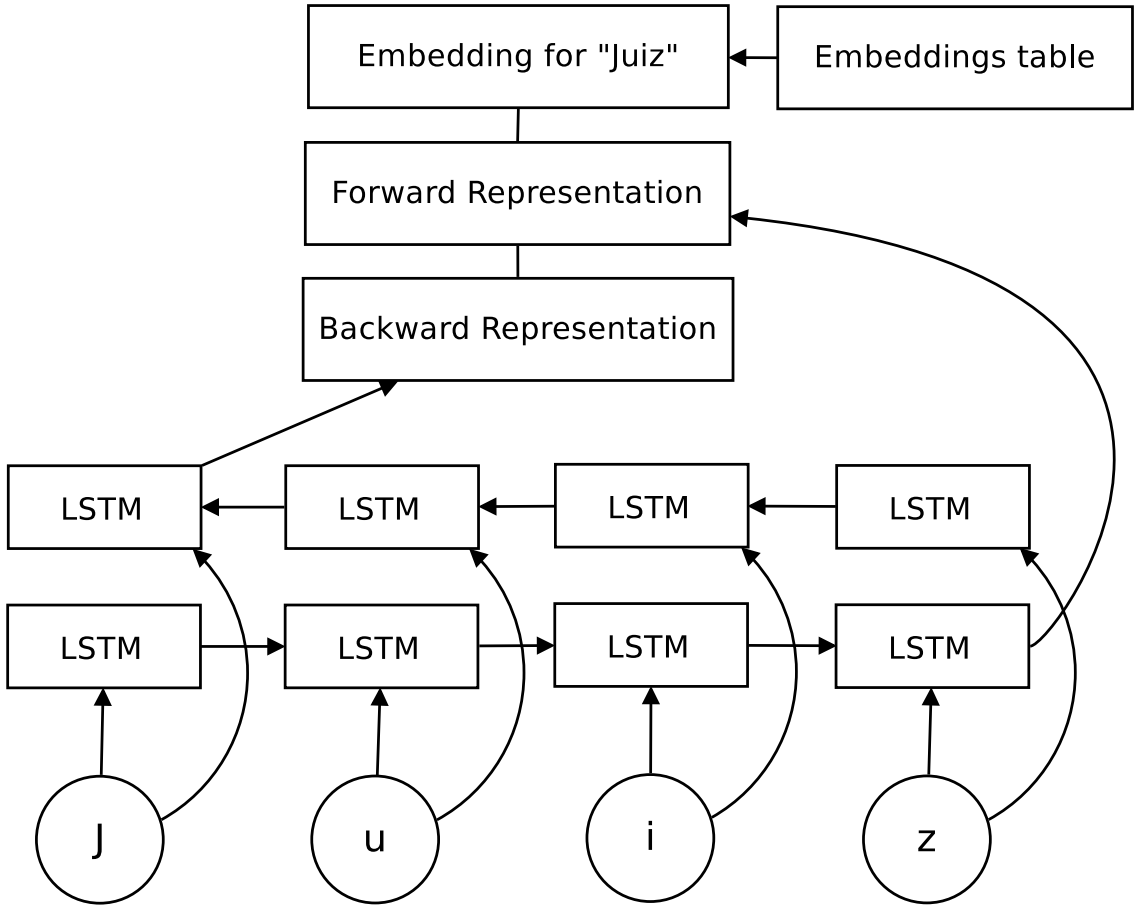
Table: LSTM-CRF results for each language.

Dataset	$\mathbf{F_1}$
English CoNLL-2003 dataset	90.94
German CoNLL-2003 dataset	78.76
Dutch CoNLL-2002 dataset	81.74
Spanish CoNLL-2002 dataset	85.75

The baseline model: LSTM-CRF

- The model input is a sequence of word vectors constructed from the concatenation of both word and character level embeddings.
- For the word-level vectors we used 300 dimensional GloVe [Pennington, J. et al, 2014] word embeddings pre-trained on a multi-genre corpus of Brazilian and European Portuguese texts[Hartmann, N. et al, 2017].
- The character-level embeddings are initialized at random values and fed to a bidirectional LSTM layer, whose output is then concatenated with the word embeddings.

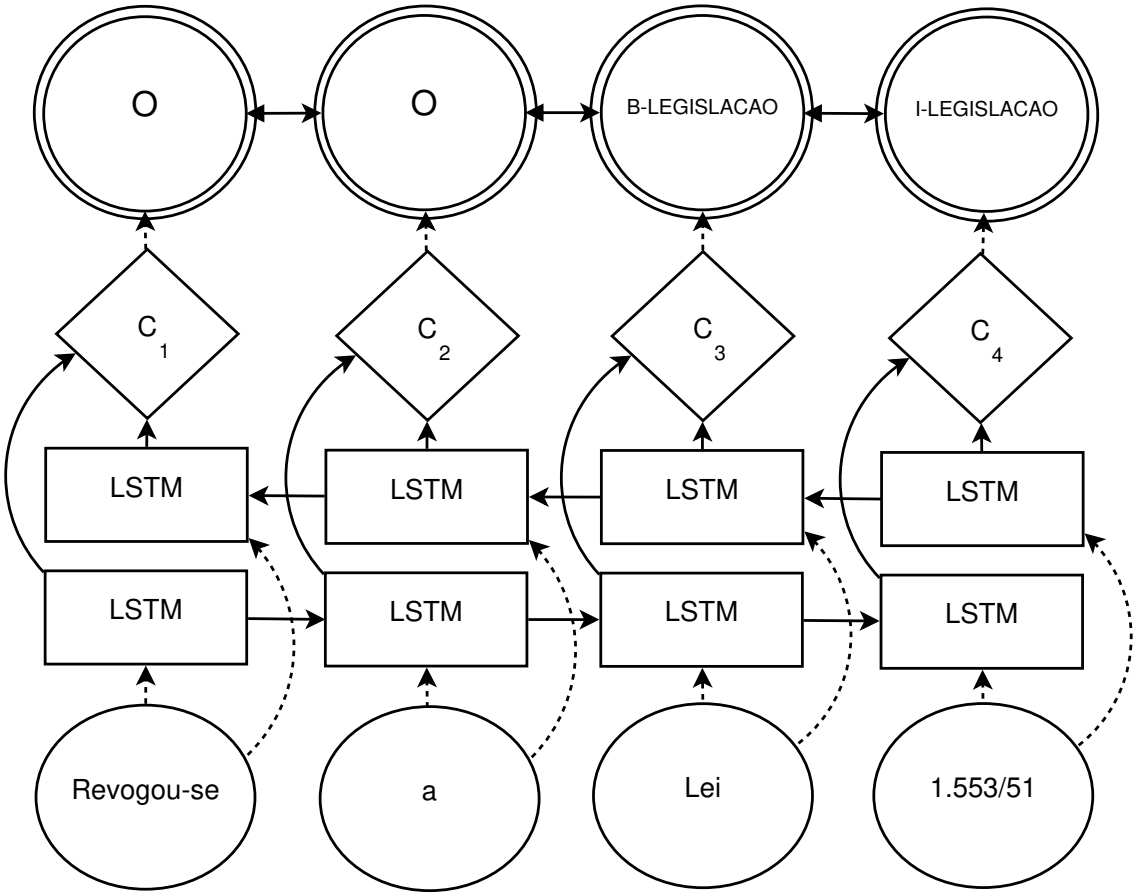
Model input


Embedding for "Juiz"
Embeddings table
Forward Representation
Backward Representation
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
J
u
i
z

The baseline model: LSTM-CRF

- The word vectors are fed to another bidirectional LSTM layer, resulting in a context vector for each word.
- A CRF layer then takes into account the context vectors and the neighboring tags to yield the predictions.
- To reduce overfitting and improve the generalization capabilities of the model, dropout masks are applied to the output of the LSTM layers.

Model Architecture



Experiments

Experiments

- The model was first trained on the Paramopama [Mendonça Jr., C.A.E. et al, 2015] corpus to evaluate if it could achieve state-of-the-art performance on a Portuguese dataset, then retrained from scratch on LeNER-Br.
- The preprocessing steps applied were lowercasing the words and replacing every digit with a zero. Such steps match the preprocessing of the pre-trained word embeddings.
  - Since the character-level representation preserves capitalization, that information is not lost.
- We tuned hyper-parameters on the development set.

Hyper-parameters

Table: 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

Results

Results

Table: Results on Paramopama Test Set 1 (10% of the WikiNER [Nothman, J. et al, 2015]).

	ParamopamaWNN[1]			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%**

[1] [Mendonça Jr., C.A.E. et al, 2016]

Results

Table: Results on Paramopama Test Set 2 (HAREM [Santos, D., Cardoso, N, 2006]).

	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%**

Results

Table: Results on LeNER-Br for token classification.

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%

Results

Table: Results on LeNER-Br for entity classification.

Entity	Precision	Recall	F_1
Person	85.58%	78.97%	82.14%
Location	69.77%	63.83%	66.67%
Organization	88.30%	82.83%	85.48%
Time	91.30%	87.50%	89.36%
Legislation	93.93%	94.18%	94.06%
Legal cases	79.29%	84.86%	81.98%
Overall	87.98%	85.29%	86.61%

Results

- The LSTM-CRF network outperforms ParamopamaWNN on both test sets.
- The results achieved on LeNER-Br are comparable to the ones achieved on Paramopama.

Results

- On the other hand, location score was noticeably lower.
  - Location entities are rare, representing 0.61% and 0.28% of the words in the train and test sets respectively.
  - Location entities are easily mislabeled as person or organization entities:
    - instead of identifying "avenida José Faria da Rocha" as a location, the model classified "José Faria da Rocha" as a person.

Results

Entities

- Person
- Time
- Organization
- Location
- Legislation
- Legal cases

Nesse sentido , rememoro manifestação do Ministro Cezar Peluso , ao julgamento do MI 822/DF : " Dispõe o § 6º do art. 57 da Lei nº 8.213, de 24 de julho de 1991, que a aposentadoria especial será custeada pela contribuição prevista no inciso II do art . 22 da Lei nº 8.212, de 24 de julho de 1991, que, por sua vez, estabelece uma contribuição social devida pela empresa na qual trabalhadores são expostos a riscos ambientais.

Results

Entities

- Person
- Time
- Organization
- Location
- Legislation
- Legal cases

Consignou-se, por exemplo, na ementa do Mandado de Injunção 592-AgR, rel. Min. Marco Aurélio, DJ de 30.04.2004, que descabia confundir preceito constitucional assegurador de um certo direito com a autorização para o legislador, em opção político-legislativa, criar exceções à regra de contagem do tempo de serviço, presentes as peculiaridades da atividade.

Results

Entities

- Person
- Time
- Organization
- Location
- Legislation
- Legal cases

(Rcl 6873/SP, rel. Min. Menezes Direito, decisão monocrática publicada no DJe divulgado em 5/11/2008, págs. 111/112)

- "Direito" should be tagged as a person.

Results

Entities

- Person
- Time
- Organization
- Location
- Legislation
- Legal cases

Vistos, relatados e discutidos estes autos de Agravo de Instrumento em Recurso de Revista n.º TST-AIRR-702/2002-070-01-40.3, em que é Agravante EMPRESA GERENCIAL DE PROJETOS NAVAIS - EMGEPRON e Agravados FERNANDO CHAVES e SOLUTION FIBER DO BRASIL LTDA.

- "Agravo de instrumento em" should be tagged as legal case;
- "EMGEPRON" should be tagged as organization.

Conclusion

Conclusion

- LeNER-BR is a Portuguese language dataset for named entity recognition applied to the legal domain.
- A state-of-the-art model trained on the dataset was able to achieve good performance - average F_1 score of 92.53% for token classification and 86.61% for entity classification.
- The dataset and source code is available at `https://cic.unb.br/~teodecampos/LeNER-Br/`
- Future work:
  - Expansion of dataset (legal documents from different courts and other kinds of legislation
  - Train word embeddings on a large corpus of legal documents and find out how domain-specific word embeddings affect the performance.

References I

Lample, Guillaume and Ballesteros, Miguel and Subramanian, Sandeep and Kawakami, Kazuya and Dyer, Chris (2012).

Neural architectures for named entity recognition.

In: Proceedings of NAACL-HLT, pp. 260–270. Association for Computational Linguistics (ACL), San Diego, 12–17 June 2016. https://arxiv.org/abs/1603.01360

Pennington, J., Socher, R., Manning, C. (2014)

GloVe: global vectors for word representation.

In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

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 Computacão, Uberlândia, MG, Brazil, 2–5 October 2017. https://arxiv.org/abs/1708.06025

References II

Mendonçaa 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).

Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.

Learning multilingual named entity recognition from Wikipedia.

Artif. Intell. 194, 151–175 (2013).

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 Inteligência Artificial e Computacional (ENIAC). SBC (2016).

References III

Santos, D., Cardoso, N.

A golden resource for named entity recognition in Portuguese.

In: Vieira, R., Quaresma, P., Nunes, M.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 69–79. Springer, Heidelberg (2006). https://doi.org/10.1007/11751984 8.