Portuguese Named Entity Recognition using LSTM-CRF

Pedro Vitor Quinta de Castro
Nádia Félix Felipe da Silva
Anderson da Silva Soares

Institute of Computing
Federal University of Goias (UFG), Brazil

September, 26 2018

{pedrovitorquinta,nadia,anderson}@inf.ufg.br

Agenda

Introduction

- **Concept:**
  - Automatic **Identification** and **Classification** of proper nouns in a text
  - Entity Extraction, NER, NERC Maynard et al. [2017]
  - Prerequisite for **Relation Extraction** and **Entity Linking**

Introduction

- **Sample annotation from HAREM dataset**

```
Sou O
Professor   B-PESSOA
Associado   I-PESSOA
do  O
Departamento    B-ORGANIZACAO
de  I-ORGANIZACAO
Matemática  I-ORGANIZACAO
da  I-ORGANIZACAO
Universidade    I-ORGANIZACAO
de  I-ORGANIZACAO
Lisboa  I-ORGANIZACAO
e   O
membro  O
de  O
o   O
Centro  B-ORGANIZACAO
de  I-ORGANIZACAO
Matemática  I-ORGANIZACAO
e   I-ORGANIZACAO
Aplicações  I-ORGANIZACAO
Fundamentais    I-ORGANIZACAO
-   O
CMAF    B-ORGANIZACAO
.   O
```

Related Work

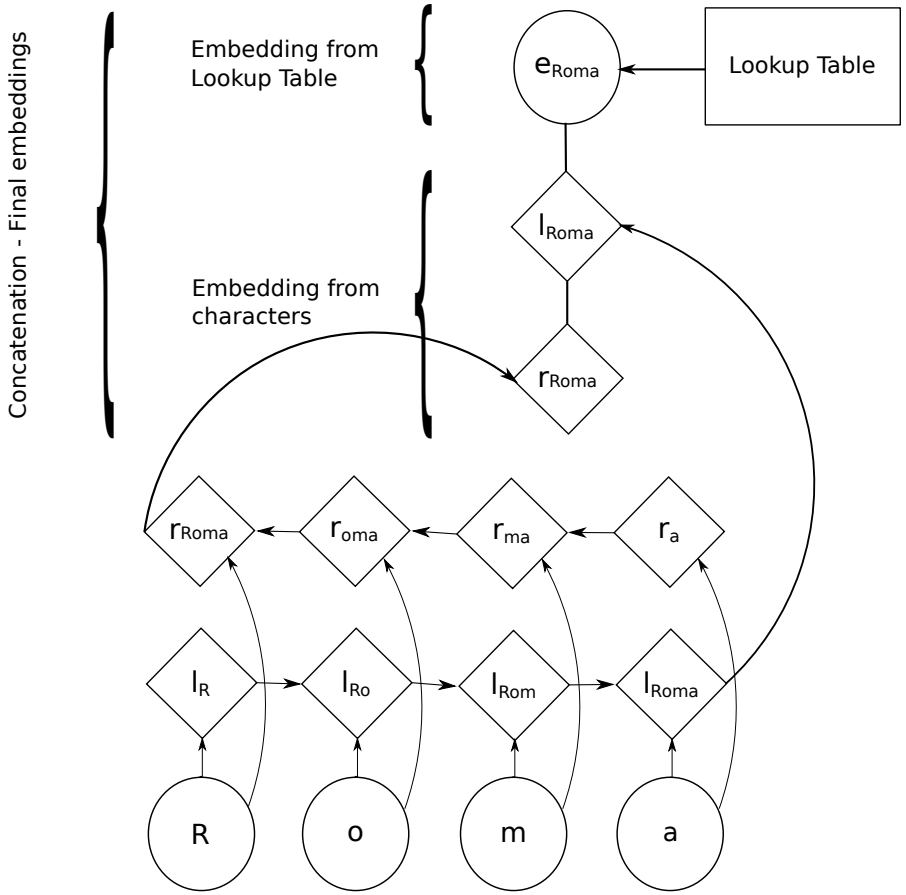
[1] Portuguese HAREM total scenario
[2] Portuguese HAREM selective scenario
[3] Spanish CoNLL-2003
[4] English CoNLL-2002

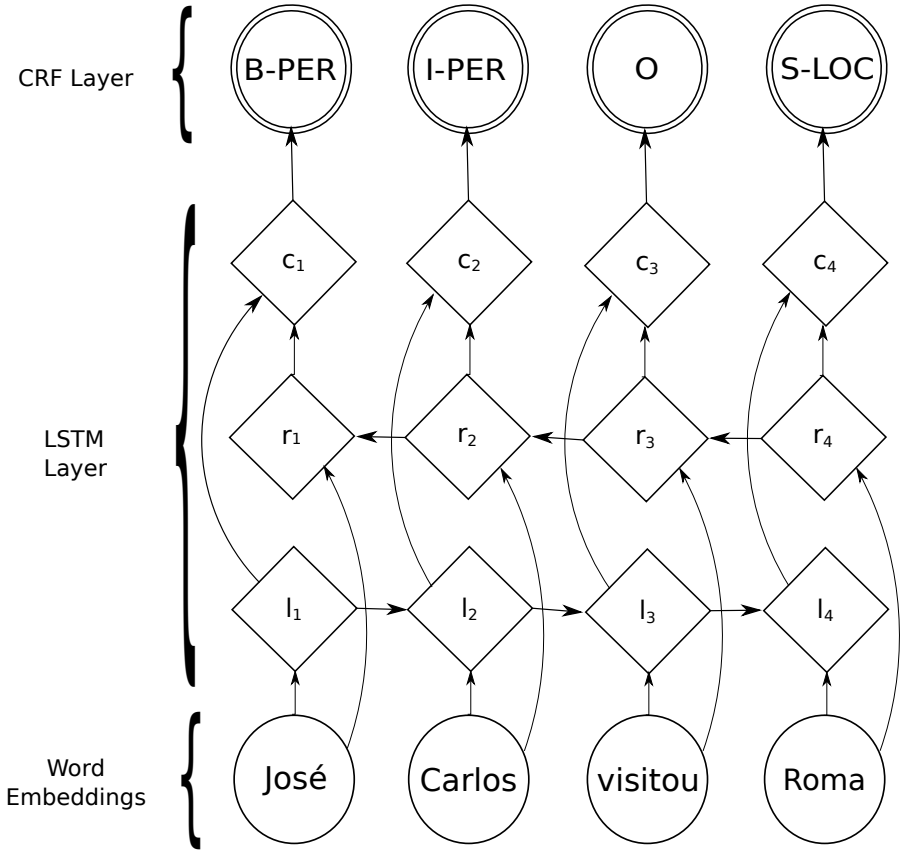
Author	Architecture	Language	Evaluation	F1 Score
do Amaral and Vieira [2014]	CRF	Portuguese	SAHARA	57.92%[1]
dos Santos and Guimarães [2015]	CharWNN	Portuguese Spanish	CoNLL	65.41%[1], 71.23%[2] 82.21%[3]
Lample et al. [2016]	LSTM-CRF	English Spanish Dutch German	CoNLL	85.75%[3] 90.94%[4]

LSTM-CRF Architecture

- **Intuition behind the architecture**
  - Sequence labeling is based on contextual information: neighboring words and how they are related
  - In order to consider a token as part of a name, some evidences must be considered:
    - **Orthographic**: Shape of the words, what do they look like (prefixes, suffixes, capitalization)
    - **Distributional**: Location of the words, how are their surroundings (morphological, syntactic and semantic information)

LSTM-CRF Architecture





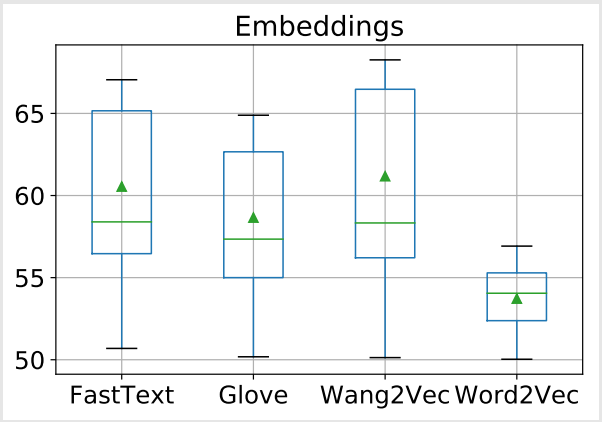
Experimental Setup

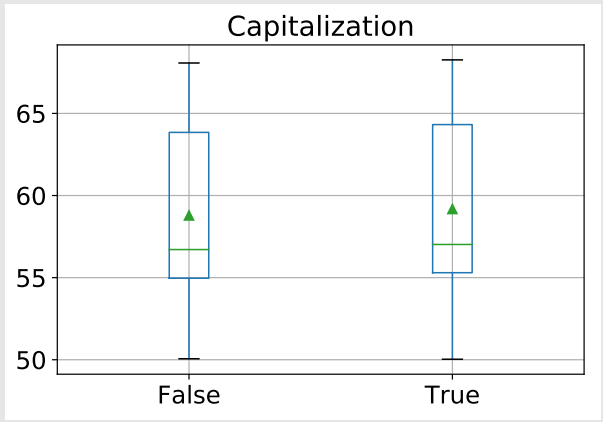
- **Hyperparameters tuning**
  - Tagging Schemes
    - **IOB2** and **IOBES**
  - Word Embeddings [Hartmann et al. 2017]
    - 100 dimensions, trained with skip-gram
    - **FastText**, **Glove**, **Wang2Vec** and **Word2Vec**
  - LSTMs Dimension
    - 25 and 50 hidden units for character-level LSTMs
    - 100 and 200 hidden units for word-level LSTMs
  - Capitalization
    - If enabled, adds an additional feature to the model according to the capitalization of the token
  - Normalization
    - If enabled, each word from the corpus is converted to lowercase prior to looking up their embeddings

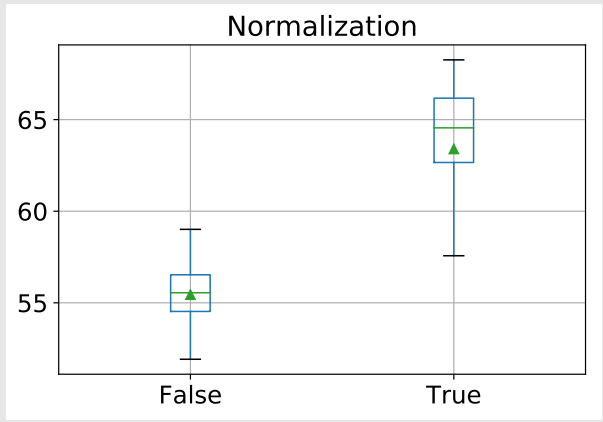
Experimental Setup

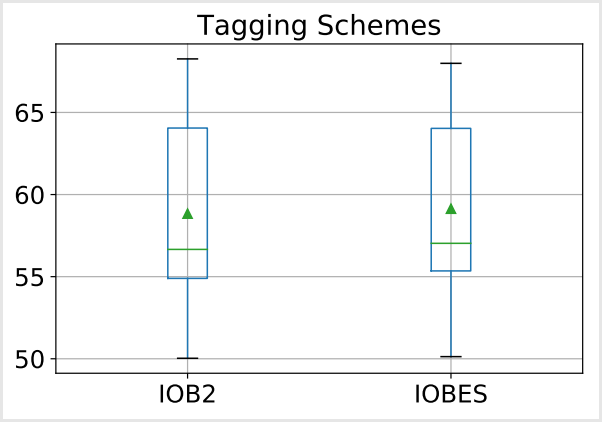
- **Training**
  - Using HAREM I for training set and MiniHAREM for test set
  - Backpropagation and Optimization with SGD, batch size 1
  - Learning rate 0.01
  - Gradient clipping 5.0
- **Evaluation**
  - Evaluation of 128 combinations of the hyperparameters' values (Tagging Schemes, Word Embeddings, LSTMs Dimension, Capitalization and Normalization)
  - Each combination was run 10 times, for 5 epochs
  - Best combinations were run 10 times, for 100 epochs

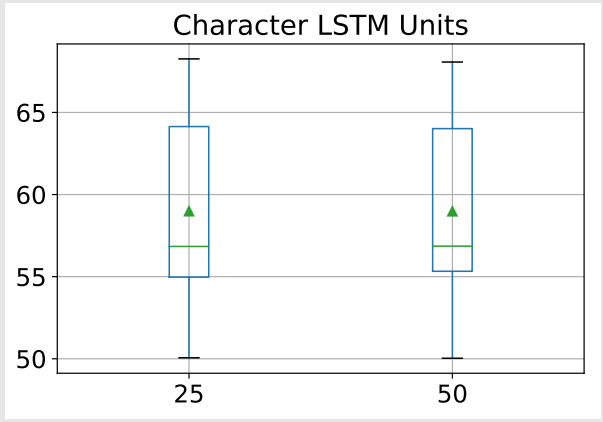
Results

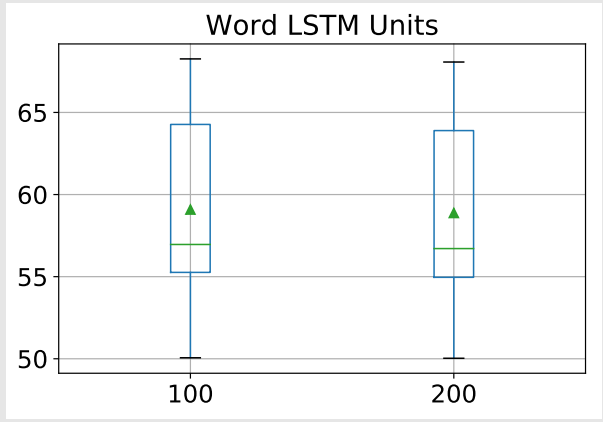












Each boxplot depicts the data related to the F1 score obtained for all 1280 executions, grouped by each type of parameter. Best combination with this criteria: ***Wang2Vec***, ***IOBES***, ***Normalization***, ***Capitalization***, ***25 hidden units for Char LSTM*** and ***100 hidden units for Word LSTM***

Results

Top 10 mean F1 scores from the different 128 evaluated scenarios

Embedding	Tagging Scheme	Normalization	Capitalization	Char LSTM Dimension	Word LSTM Dimension	Mean F1	Max F1
Wang2Vec	IOB2	True	False	50	100	66.90	67.69
Wang2Vec	IOB2	True	True	25	100	66.89	68.26
Wang2Vec	IOB2	True	False	50	200	66.82	68.07
Wang2Vec	IOBES	True	True	25	200	66.79	67.99
Wang2Vec	IOB2	True	True	50	100	66.67	67.64
Wang2Vec	IOB2	True	True	50	200	66.60	67.22
Wang2Vec	IOBES	True	True	25	100	66.55	67.87
Wang2Vec	IOBES	True	True	50	100	66.45	67.79
Wang2Vec	IOBES	True	False	50	100	66.37	67.58
Wang2Vec	IOBES	True	True	50	200	66.33	66.78

Best combination would be: ***Wang2Vec***, ***IOB2***, ***Normalization***, ***No Capitalization***, ***50 hidden units for Char LSTM*** and ***100 hidden units for Word LSTM***

Results

Comparison with the state-of-the-art for the HAREM I corpus

Architecture	Total Scenario			Selective Scenario		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
CharWNN	67.16%	63.74%	65.41%	73.98%	68.68%	71.23%
LSTM-CRF *	**72.78%**	68.03%	70.33%	78.26%	**74.39%**	**76.27%**
LSTM-CRF **	72.45%	**68.63%**	**70.48%**	**78.39%**	73.83%	76.03%

* Model trained using each of the best individual parameters

Wang2Vec, IOBES, Normalization, Capitalization, 25 hidden units for Char LSTM and 100 hidden units for Word LSTM

** Model trained using the best combination of parameters

Wang2Vec, IOB2, Normalization, No Capitalization, 50 hidden units for Char LSTM and 100 hidden units for Word LSTM

Conclusions

- **We improved existing benchmark of Portuguese NER for HAREM corpora by experimenting with the LSTM-CRF Deep Learning architecture:**
  - For the Portuguese word embeddings used:
    - Wang2Vec was the best evaluated embedding type;
    - Applying word normalization is imperative.
- **Future work:**
  - Train a NER model using corpora belonging to a specific domain, such as annotated legal texts;
  - Transfer learning with a pre-trained language model instead of word embeddings.

Acknowledgements

AVISO URGENTE®
Confiança Jurídica

Questions?

References I

do Amaral, D. O. F. and Vieira, R. (2014). NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.

dos Santos, C. N. and Guimarães, V. (2015). Boosting named entity recognition with neural character embeddings. *CoRR*, abs/1505.05008.

Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *CoRR*, abs/1708.06025.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.

Maynard, D., Bontcheva, K., and Augenstein, I. (2017). *Natural Language Processing for the Semantic Web*. Morgan and Claypool.