Bidirectional LSTM with a Context Input Window for Named Entity Recognition in Tweets

Rafael Peres
Federal University of Rio de Janeiro, Brazil
rafaelperes@ufrj.br

Diego Esteves
University of Bonn, Germany
SDA Research
esteves@cs.uni-bonn.de

Gaurav Maheshwari
University of Bonn, Germany
SDA Research
gaurav.maheshwari@uni-bonn.de

ABSTRACT

Lately, with the increasing popularity of social media technologies, applying natural language processing for mining information in tweets has posed itself as a challenging task and has attracted significant research efforts. In contrast with the news text and others formal content, tweets pose a number of new challenges, due to their short and noisy nature. Thus, over the past decade, different Named Entity Recognition (NER) architectures have been proposed to solve this problem. However, most of them are based on handcrafted-features and restricted to a particular domain, which imposes a natural barrier to generalize over different contexts. In this sense, despite the long line of work in NER on formal domains, there are no studies in NER for tweets in Portuguese (despite 17.97 million monthly active users). To bridge this gap, we present a new gold-standard corpus of tweets annotated for Person, Location, and Organization (PLO). Additionally, we also perform multiple NER experiments using a variety of Long Short-Term Memory (LSTM) based models without resorting to any handcrafted rules. Our approach with a centered context input window of word embeddings yields 52.78 F1 score, 38.68% higher compared to a state of the art baseline system.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; *Information extraction*;

KEYWORDS

Deep learning, neural networks, machine learning, informal text, named entity recognition, natural language processing

ACM Reference Format:
Rafael Peres, Diego Esteves, and Gaurav Maheshwari. 2017. Bidirectional LSTM with a Context Input Window for Named Entity Recognition in Tweets. In *Proceedings of K-CAP 2017: Knowledge Capture Conference (K-CAP 2017).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3148011.3154478

K-CAP 2017, December 4–6, 2017, Austin, TX, USA

ACM ISBN 978-1-4503-5553-7/17/12…$15.00
https://doi.org/10.1145/3148011.3154478

1 INTRODUCTION

In many natural language processing (NLP) information extraction (IE) pipelines the task of Named Entity Recognition (NER) is a relevant step. It involves the identification of proper names in texts and their classification into a set of predefined categories of interest. Despite it seems to be a simple task to perform, it poses a number of challenges notably in noisy data sets such as Twitter messages, also known as *tweets*. This is so prominent that it is still considered as an emerging research field, even for English corpora [4]. Although recent efforts, we still face limitations in correctly identifying and classifying entities. State-of-the-art NER systems in formal *English* texts currently have about 85-90% accuracy - such as articles (CoNLL03 shared task dataset). However, NER applied in short, noisy and informal texts still perform poorly (about 30-50% accuracy), mainly due to the lack of implicit linguistic formalism (e.g. proper punctuation, spelling, spacing, formatting, unorthodox capitalisation, emoticons, abbreviations and *hashtags*) presented [4]. Additionally, the lack of external knowledge resources is an important gap in the process regardless of writing style. To face these problems, prior research had primarily focused on microblog-specific IE techniques [17].

In this paper, we look at the NER task in noisy and short texts such as tweets following recent advances using neural architectures, showing better results than traditional NER approaches. Furthermore, in order to increase available resources to the community, a *Portuguese* data set for NER is annotated and released[1]. *Portuguese* is one of the most spoken languages in the World and has increasing popularity on Twitter, with 17.97 million monthly active users[2]. In particular, we make the following contributions. (1) we release a new corpus of Twitter messages annotated for both Person (PER), Location (LOC) and Organization (ORG) in one of the most spoken languages in the World; (2) we analyze a variety of neural network (NN) models and features to NER task on noisy texts based on state-of-the-art architectures [12]. (3) we present the first benchmark report of a NER architecture on noisy text from social media using neural networks; (4) we show similar results when compared with other languages (*English* and *French*) confirming neural networks as a better choice to noisy and short texts opposing to traditional NER architectures.

The rest of this paper is organized as follows: Section 2 presents previous works on NER task. Section 3 describes the corpus annotation process and its statistics. Section 4 describes the models used to the proposed task. In Section 5, the experiments conducted over

[1] http://bit.ly/2xJ8Muw
[2] statistics from May 2016: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

the presented data set and the analysis of the results are presented. Finally, in Section 6, we present conclusions about this work.

2 RELATED WORK

In recent years, named entity recognition in natural language texts has been addressed by a number of different approaches. Most of existing frameworks heavily rely on look-up strategies and encoded rules, as well as standard local features, such as part-of-speech tags, previous/next words, shapes and REGEX expressions [16]. This architecture, as expected, reports a steep fall in performance with noisy data, such as Tweets data sets. A major reason is that they rely heavily on handcrafted features and domain-specific knowledge. More recently, studies have successfully proposed neural architectures to solve this problem [11, 12, 15]. Chiu and Nichols [1] proposed a neural network architecture that automatically detects word and character-level features using a hybrid Bidirectional Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN). With respect to *Portuguese* language Santos and Guimarães [5] successfully employed CNN's to extract character-level features. Thus, these models work without resorting to any language-specific knowledge or resources such as *gazetteers*. They, however, focused on *newswire* domain to improve current state-of-the-art systems and not on the social media domain, in which they are naturally much harder to outperform due to the issues observed. In this context, some approaches have been proposed, but they are mostly still in development and are often not freely available [4]. Recently, Limsopatham and Collier [13] employed a Bidirectional LSTM model for *English* tweets reaching the first position in a shared task with F1 52.41 [19]. In addition to that, for *French*, neural network models have also achieved leading positions [14]. Finally, Esteves et al. proposed a different NER approach, applying embedding knowledge derived from heuristics obtained from images and texts clustered in a token level perspective. This approach reached state-of-art results using tweets in a popular *English* data set for microblogs [7].

3 CORPUS DESCRIPTION

We constructed a NER corpus of Twitter messages in *Portuguese* (`pt-br`), following Finin et al. [8] annotation guidelines for the entities PER, LOC and ORG and applying the *CoNLL03* shared task data format [20]. We collected general data using the Twitter API[3] with location and language adjusted to *Portuguese*. Our Twitter data set comprises daily tweets between 17^{th} and 25^{th} April 2017. In order to minimize poor temporal diversity and the entity drift phenomenon [3], for each hour we collected 200 random tweets, with a total of 38400 tweets collected. From this sample, we randomly choose 4000 tweets to annotate. The BRAT tool[4] was used for annotation (Figure 1). The final corpus has a total of 3.968 tweets after some data cleaning operations like deleting non-Portuguese tweets. The total number of tokens is 44951 and the average number of tokens per tweet is about 11.32. The training and development data for our corpus was randomly split in 80% for training data and 20% for test data. The final corpus descriptive statistics is showed in 1.

[3]https://dev.twitter.com/rest/public
[4]http://brat.nlplab.org/

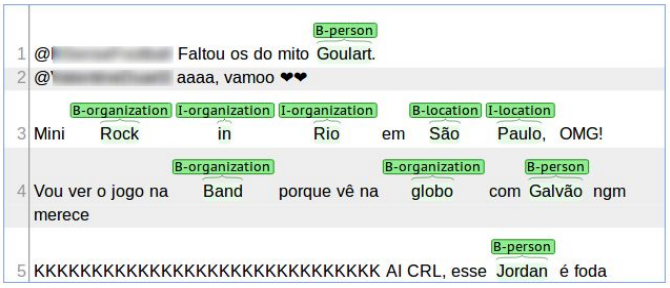

Figure 1: Examples of annotated tweets used

4 FEATURES AND MODEL

In this section, we describe the components and features related to our implemented NN model. We introduce the components as follows:

4.1 Word Embeddings

We used the openly available *GloVe 100-dimensional embeddings*[5] trained on 27 billion of tokens from Twitter. In addition, in order to test the effectiveness of the pre-trained word embeddings in noisy and short texts, we experimented with randomly initialized embeddings with 100 dimensions as proposed by Ma and Hovy using formal texts domain [15]. For the words that do not exist in the pre-trained embeddings, we used a vector of random values sampled from $[-\sqrt{3/dim}, +\sqrt{3/dim}]$ where dim is the dimension of the word embeddings [9][15].

4.2 LSTM

We implemented Long Short-Term Memory (LSTM) [10] based models for the task of sequence tagging. LSTM networks are usually better at finding and exploiting long-range dependencies in data, an important characteristic in the NER task. This behavior is present mainly because their hidden layer updates are replaced by purpose-built memory cells. We used the following implementation:

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$
$$h_t = o_t \circ \sigma_h(c_t) \quad (5)$$

Where σ is the logistic sigmoid function, $\circ$ is the element-wise product and i, f, o and c are the input gate, forget gate, output gate and cell vectors, all of which are the same size as the hidden vector h. W, U and b are parameter matrices and vector, respectively.

4.3 Bidirectional LSTM (BI-LSTM)

NER is a *sequence tagging* task, therefore for an input feature in a given time *t*, accessing both past and future information is of utmost relevance. A known limitation of LSTM models in this context is that the hidden state *ht* takes information only from past, knowing nothing about the future. To bridge this gap, Bidirectional LSTMs

[5]https://nlp.stanford.edu/projects/glove/

Table 1: Our Twitter data set descriptive statistics. In addition, a summary of the English and French data set.

	Person	Location	Organization	Tokens	Tweets
Train	409	162	174	35800	3175
Test	107	33	50	9151	793
Total	**516**	**195**	**224**	**44951**	**3968**
Ritter et al. 2011 (English)	454	380	387	34000	2400
Lopez et al. 2017 (French)	1506	1264	736	Not reported	6885

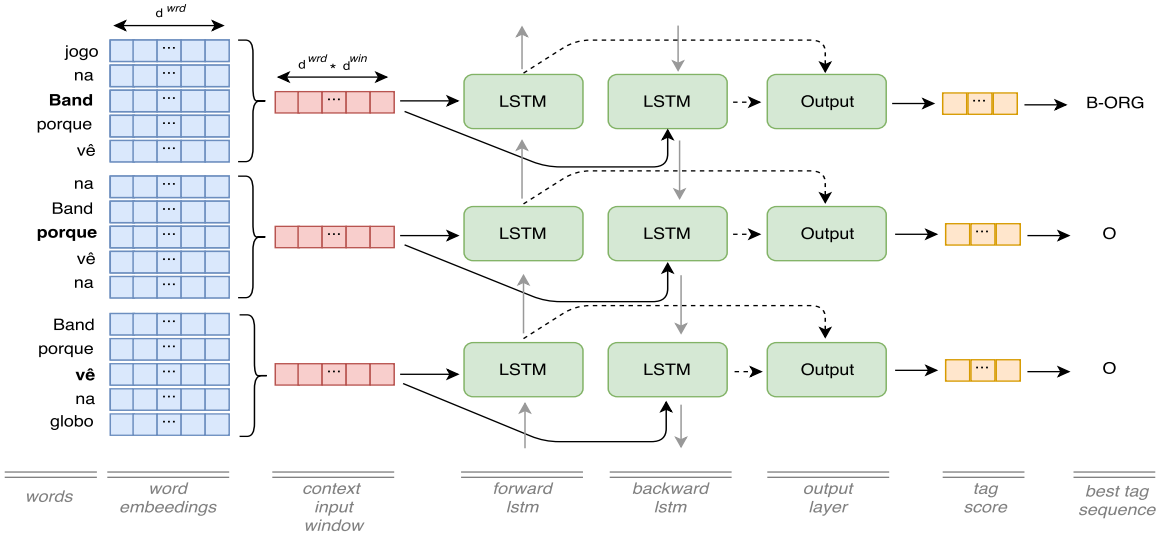

Figure 2: The BI-LSTM for tagging named entities. For each word, vectors in a window centered context are fed to the BI-LSTM network and then output a *Tag sequence*. Dashed arrows indicate dropout layers applied.

(BI-LSTM) have been shown as an efficient solution [1, 12, 13]. The main idea behind BI-LSTM is to make use of both past and future features, via *backward* and *forward* states, respectively, for a specific time frame and concatenate them to form the final output.

4.4 Proposed Model

Our model is encouraged by the seminal work of Collobert et al. [2]. From this work, we adapted the idea of input *window approach*. Our neural network takes as input a fixed-sized window of word embeddings centered around one target word *w*, using previous and following words in order to capture the context in *short, informal and noisy texts* such as in our study: *tweets*. In addition to that, as revealed by recent state-of-the-art results in newswire data sets [1, 11, 12, 15], our approach used a BI-LSTM network to receive the *context input window* presented. The final architecture is depicted in Figure 2.

5 EXPERIMENTS AND RESULTS

We first conducted experiments to study the effectiveness of each component of our NN architecture by *ablation studies* [15]. We compared our results with the baseline system Stanford NER [18], a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models, trained and evaluated on our data set in order to show our approach against more traditional systems.

5.1 Parameter Optimization

Parameter optimization is performed with Adadelta [21]. We used a mini-batch size of 50. According to our initial experiments, "appropriate" parameters appear at around 80 epochs based on early stopping performance on validation sets (used part of training data for validation purpose - 20%). Dropout is applied in order to mitigate over-fitting and regularization. We fixed dropout rate at 0.5 for all dropout layers through all the experiments. Table 2 summarizes the chosen hyper-parameters for all comparative experiments performed.

5.2 Results

We trained both LSTM and BI-LSTM models with the *context input window* and a learning rate of 1. In addition, we performed two ways to initialize the word embeddings: *Random* and *Glove Twitter*. For each experiment conducted, we used identical feature sets, therefore different outcomes presented are only due to different networks. We report models performance on test data sets in Table

Table 2: Hyper-parameters for all experiments

Layer	Hyper-parameters	Number
LSTM	State size	150
LSTM	Initial state	0
LSTM	Peepholes	No
	Epochs	80
	Dropout	0.5
	Mini-batch size	50

Table 3: Performance of our proposed model (Best F1 score), together with a baseline system and results of similar models in other Languages

Model	Precision	Recall	F1
Stanford NER (pt-br)	**68.92**	26.29	38.06
LSTM + word (Random)	21.13	28.87	24.40
LSTM + word (Glove)	49.75	50.52	50.13
LSTM + word (Glove) + window	51.49	**53.61**	52.53
BI-LSTM + word (Random)	22.22	30.93	25.86
BI-LSTM + word (Glove)	50.75	52.58	51.65
BI-LSTM + word (Glove) + window	57.23	48.97	**52.78**
BI-LSTM (English Shared Task)	46.07	60.77	52.41
BI-LSTM (French Shared Task)	58.95	46.83	52.19

Table 4: Best model performance

Best Model	Precision	Recall	F1
Person (PER)	67.68	62.04	64.73
Location (LOC)	32.00	22.86	26.67
Organization (ORG)	47.62	39.22	43.01
Overall	**57.23**	**48.97**	**52.78**

3 using *CoNLL03 shared task* evaluation script[6]. For reproducibility purposes, the experiment configurations were exported based on the MEX Vocabulary [6]. According to the results showed in Table 3, the major improvement uses pre-trained embeddings as input features. Regarding the models, BI-LSTM with pre-trained embeddings and a *context window input* with size 5 achieves the best $F1$ performance with a score of 52.78, 38.68% higher than the proposed baseline. The detailed results regarding the entities Person, Location and Organization of this model are described in Table 4.

6 CONCLUSIONS

In this paper, we described and presented our *Portuguese* (pt-br) NER twitter corpus. We also performed an evaluation for multiple NER experiments using a variety of recent state-of-the-art NN architectures. We systematically compared the performance of LSTM networks based models for sequence tagging and showed that without resorting to hand-crafted features, our proposed model is effective for the Twitter NER tasks and achieves better performance over a traditional system. We showed that a *context input window* approach in both LSTM and BI-LSTM models also boosted results. Furthermore, we presented a similar result compared to other languages, such as *English* and *French*. It shows how relevant neural networks approaches are and how they can be highly effective for NER in short and noisy texts, shedding light for future studies.

[6] http://www.cnts.ua.ac.be/conll2002/ner/bin/conlleval.txt

REFERENCES

[1] Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional LSTM-CNNs. *arXiv preprint arXiv:1511.08308* (2015).

[2] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* (2011).

[3] Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *COLING*.

[4] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* 51, 2 (2015), 32–49.

[5] Cícero Nogueira dos Santos and Victor Guimarães. 2015. Boosting Named Entity Recognition with Neural Character Embeddings. *CoRR* abs/1505.05008 (2015).

[6] Diego Esteves, Diego Moussallem, Ciro Baron Neto, Tommaso Soru, Ricardo Usbeck, Markus Ackermann, and Jens Lehmann. 2015. MEX vocabulary: a lightweight interchange format for machine learning experiments. In *Proceedings of the 11th International Conference on Semantic Systems*. ACM, 169–176.

[7] Diego Esteves, Rafael Peres, Jens Lehmann, and Giulio Napolitano. 2017. Named Entity Recognition in Twitter using Images and Text. *The International Conference on Web Engineering* (2017).

[8] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 80–88.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015).

[12] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

[13] Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for Named Entity Recognition in Twitter Messages.

[14] Cédric Lopez, Ioannis Partalas, Georgios Balikas, Nadia Derbas, Amélie Martin, Coralie Reutenauer, Frédérique Segond, and Massih-Reza Amini. 2017. CAp 2017 challenge: Twitter Named Entity Recognition. *CoRR* abs/1707.07568 (2017). http://arxiv.org/abs/1707.07568

[15] Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[16] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

[17] Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1524–1534.

[18] Jenny Rose Finkel, Trond Grenager, and Christoper Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. (01 2005).

[19] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 Named Entity Recognition Shared Task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*.

[20] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. 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 2003 - Volume 4 (CONLL '03)*.

[21] Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR* abs/1212.5701 (2012). http://dblp.uni-trier.de/db/journals/corr/corr1212.html#abs-1212-5701