A New Entity Extraction Model Based on Journalistic Brazilian Portuguese Language to Enhance Named Entity Recognition

Rogerio de Aquino Silva, Luana da Silva, Moisés Lima Dutra, and Gustavo Medeiros de Araujo(✉)

Engineering and Data Science Lab, Federal University of Santa Catarina, Florianópolis, Brazil
{rogerio.aquino,silva.luana}@posgrad.ufsc.br,
{moises.dutra,gustavo.araujo}@ufsc.br

Abstract. Named Entity Recognition (NER) plays an important role on broad natural language processing applicability. According to the literature, the NER process applied to the English language reaches around 90% of accuracy. However, when applied to Portuguese, this accuracy is at most 83.38%. A wide range of algorithms based on LSTM (Long-Short Term Memory) architecture has being proposed to enhance the NER accuracy. However, a key component to a successful information extraction is the corpora used for NER training. In order to improve the NER in Portuguese language, this paper proposes a methodology for training text corpus based on Portuguese-language journalistic corpora. The Journalistic language has the best adherence to the contemporaneity of the language, since it preserves features such as objectivity, simplicity, impartiality, and is a reference of transmitting the information without ambiguity. The proposed methodology provides a model to extract entities and assess the obtained results with the use of Recurrent Neural Network architectures. At the best of our knowledge, with the proposed methodology, the NER task applied to the Portuguese language overcomes the average accuracy found in the literature, increased from 83.38% to 85.64%. Moreover, the use of this methodology could decrease the computational costs related to the NER processing tasks.

Keywords: Natural Language Processing · Name entity recognition · Entity extraction model · Brazilian Portuguese corpus · Recurrent Neural Networks

1 Introduction

Information retrieval (IR) has emerged from efforts on facilitating large-scale data manipulation [8]. Efforts to IR development still face major challenges when

R. Mugnaini (Ed.): DIONE 2020, LNICST 319, pp. 53–63, 2020.
https://doi.org/10.1007/978-3-030-50072-6_5

it comes to Natural Language Processing (NLP). The extraction of information from Portuguese texts is still an open field of investigation. It is a consequence of the weak results of NLP models for Portuguese language [2].

In addition to the challenge of applying NLP for Portuguese, there is a growing scale of data generated by the increasing number of internet users [7]. It is estimated that at every second thousands of data are generated in a variety of formats such as images, text, videos, and audios.

A study conducted by Data Management Association (Dama) found there are currently over 500 quadrillion of Megabytes of data stored in the digital universe. The study also pointed out that every two years, the production of data doubles, forecasting to reach 350 zettabytes by 2020 [9]. The large majority of the data generated are unstructured data, i.e. text written in natural language. The increasing scale of text production and the low average accuracies obtained when applying NER to Portuguese are issues that can be tackled with modern machine learning techniques, such as artificial neural networks, decision trees, support vector machines, and Bayesian networks.

By using such techniques, it is possible to create models with the ability to extract entities from texts in natural language. In the English language, for example, some models achieve accuracy of 92.6% when using the spaCy[1] framework, and of 91.7% with the ClearNLP[2] framework, which focus on the training of models for entity recognition.

In order for a model to achieve high performance, it is necessary to provide it with a pre-sorted dataset that possesses notations about entities and the grammatical structure. This dataset is known as textual corpus. Usually, the textual corpus is constructed by the aggregation of text excerpts, which is performed by linguists who know the language structure and its properties. However, since each corpus is created for a purpose they do not have always the same structure, so there is no pattern between them.

Moreover, [19] points out that the morphology and syntax of the Portuguese language have their own characteristics. When compared to the English language, which has fewer elements in its grammatical notation – specially in the conjugation of verbs –, those characteristics generate a more complex scenario to deal with.

Language resources are an important feature when developing computational methods to analyze and study languages [3]. In order to build accurate classifiers, there is a need for quality corpora in order to develop and evaluate classifier models [12]. The key problem in this area of research is how to build large natural language corpora enriched with morphosyntactic information [3]. There are just a few available corpora for Portuguese NER, in which the annotated datasets are usually small and/or insufficient for achieving high accuracies for Portuguese NER. Therefore, reproducing and benchmarking the results of previous works is

[1] https://spacy.io/.
[2] https://github.com/clearnlp.

not simple due to the variety of possible dataset combinations and the lack of a standardized methodology for training and evaluating [17]. According to [16], in Portuguese the accuracy is around 83.38%, when one uses the HAREM[3] corpus as a training base to be taken in CRF neural networks.

The scope of information is also a matter to be considered. According to [5], a particular community may have habits regarding the use of information, i.e. how to perform its searches and how to organize its new knowledge. These habits affect not only how the text is written but also how some terms are syntactically structured, even when different communities share the same language. When considering the journalistic writing, a style that has the best adherence to the contemporaneity of the language, some of the characteristics of it such as objectivity, simplicity, impartiality, and referential, use to avoid unusual terms because the information must be clearly transmitted to the reader [15]. For this reason, we chose textual corpora based on journalistic texts to create a training base for our NER approach.

Several machine learning techniques allow the creation of models capable of recognizing entities. One of the most used for this purpose is Recurrent Neural Networks (RNN) [11]. RNNs are generally applied to problems where there is a need for pattern recognition that varies over a given series [10]. However, NER is not a simple task. Several categories of named entities are written similarly and appear in similar contexts. Furthermore, the same named entity can be classified into different categories depending on the surrounding context and some entities do not appear even in large training datasets [14]. Due to the fact that reproducibility is hard when the corpus is not standardized, this work aims to create a new corpus for the Portuguese language based on journalistic texts extracted from the CETENFolha[4].

In order to test the corpus proposed in this work, we used a recurrent network variation, called Bidirectional Long-Short Term Memory (Bi-LSTM) [20]. The proposed model has been trained based on language syntax and allows new information to be added to the training, resulting in a better classification of texts in a given domain.

The remainder of the paper is structured as follows. Section 2 provides a summary of some works related to corpus building. Section 3 presents the methodology used to build and validate the proposed corpus. Section 4 presents an evaluation of the proposal. Finally, Sect. 5 provides the conclusions of this paper.

2 Related Work

A large number of proposals aiming to develop new corpora for classifiers can be found in the literature. The corpora can be based on social media data, open data from news feed, websites or even data from corporations [12]. The authors

[3] https://www.linguateca.pt/HAREM.
[4] https://www.linguateca.pt/cetenfolha.

of [6] developed a corpus for smart environments. It comprises audio and video materials, as well as robot and apartment reactions, and information taken from sensors and actuators. The data was gathered from 62 volunteers and can be used for training automated robots. As argued by the authors, this data is valuable for further in-depth analyses of people's interactions with devices, ambient intelligence and robots in everyday environments. The difference between our work and the work of [6] is that we focus on text data for providing corpora for NLP tasks.

In [4] is presented a domain specific Question-Answering Corpus (QA-Corpus) built with Portuguese tweets and news articles. While using social media, the authors could gather candidate and reliable answers to possible user questions, making the dataset more real. They used deep learning to match questions and rank candidate answers. Both this paper and [4] attempts to create corpora for Portuguese language, which lacks of good datasets/corpora in order to benchmark results. As opposed to [4], which focuses on Q&A Systems, our work focuses on NLP NER task.

A Czech attempt to create a national corpus was reported by [18]. The authors describe a project to build a larger corpus comprised of Czech texts extracted from web pages. The motivation behind this project lies on the fact that the authors believe that large corpora are essential to modern methods of computational linguistics and natural language processing. The difference between our work and [18] is that our work does not use web pages, but improves existing corpora in order to produce a unique, unified and comparable corpus.

An Arabic strive to create a corpus took place by relying on online published newspapers from different Arabic countries. The authors of [1] created the corpus for improving different researches in Information Retrieval, Machine Translation, and Arabic Language Processing, in general. As well as [1], our paper is an attempt to standardize a big and comparable corpus for Portuguese language processing in several areas of research. Besides, both works use newspapers text to build the dataset/corpus.

The authors of [13], made a corpus for a specific language domain, the legal vocabulary. The corpus contains entities such as *"TEMPO"*, *"JURISPRUDENCIA"* and *"LEGISLACAO"*. The created model used neural networks LSTM-CRF and LSTM-CNN, with an accuracy of 90.01%. On the other hand, the study [16] of 2019 use the corpus HAREM in two different scenarios. In the first, the corpus is used in its entirety, with the entities *"PESSOA"*, *"ORGANIZACAO"*, *"LOCAL"*, *"VALOR"*, *"TEMPO"*, *"ABSTRACCAO"*, *"OBRA"*, *"ACONTECIMENTO"*, *"COISA"* and *"OUTRO"*. In the second scenario, it was considered just the entities *"PESSOA"*, *"LOCAL"*, *"ORGANIZAÇÃO"*, *"DATA"*, and *"VALOR"*. For the first scenario, the accuracy obtained was 74.91%. In its turn, the second scenario obtained 83.38% accuracy. The models used a variation of a set of neural networks like Recurrent LSTMs and convolutional networks CNNs and framework flair.

3 Proposed Methodology

In this section, we present the methodology used for standardizing Portuguese corpora in order to produce a single corpus to be used by NER tasks. The methodology was developed in four modules, as it is shown in Fig. 1.

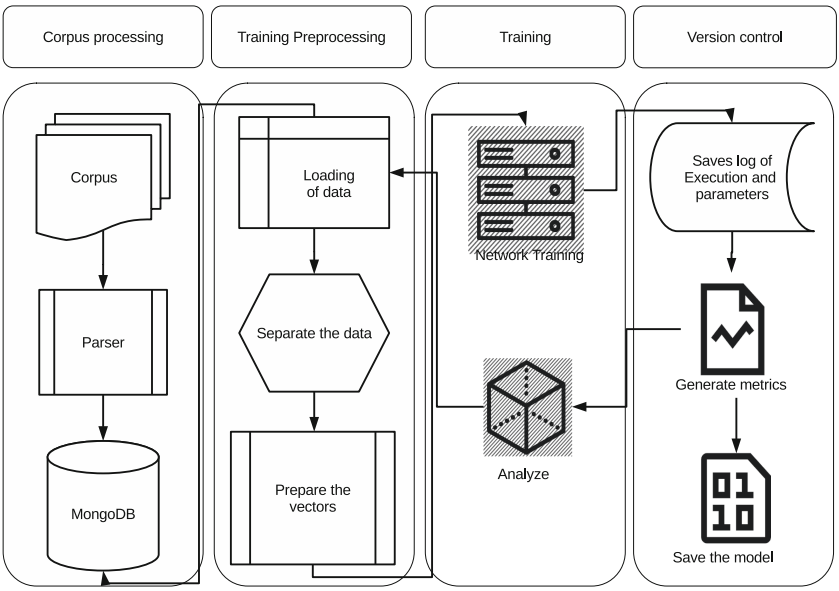

Fig. 1. Structure organization

3.1 Corpus Processing

The first step comprises the parsing of the text to check it for compliance with grammar rules. To create the base corpus, we used the CETEMPublico corpus, which possess a structure in which each line represents a token. A token can be a word, a text fragment, or a punctuation element. A token can be of four distinct types: i) infinitive verb; ii) noun; iii) POS-tag, which indicates a grammatical class; and iv) grammatical detail, which in the case of a verb, would contain the verb itself and its conjugation.

After the parsing, it is necessary to organize the data. Typically, the corpus comprises text files and these files can be massively large, which makes them expensive to process. The process of generating the training base is usually done with multiple datasets at the same time, which can make training which can make training extremely costly computationally. However, the use of a nonrelational database is proposed to mitigate this problem. A structure was created

after parsing the information, so that it is organized and inserted into a nonrelational database. This way, other project modules can easily access datasets as needed to generate pre-training.

After parsing and inserting the data into the database, it is necessary to convert the data, as annotations entered into the corpus must meet the demands of non-technology related areas and end up being confusing or not having a pattern on all lines. When training is performed using a neural network we need the data to have the same structure, because if there are incorrect notations the model runs the risk of not being able to generalize and, consequently, not predict correctly. The steps performed were: i) all tokens "PROP" POS-tag were localized, since they are proper names and it is then possible to use them to locate implicit entities; ii) next, the other annotations related to this token that contain *<inst><org><media>* xml tags were located and converted to ORG, which stands for organization entity; iii) *<hum>* tags were converted to PER (person entity), and *<civ>* to LOC (location entity). When this process is finished, the data is sent in the form of three vectors to the training preprocessing module, along with POS-Tag annotations and a list of entities.

3.2 Training Preprocessing

After the training data is received, it is necessary to collect all text excerpts from the vectors, in order to create a unique word vector that will form our final vocabulary known as "word2index". This unique vector is a dictionary containing words and their codes. The same process occurs with the POS-Tag and entity vectors.

Artificial neural networks require data to be sent numerically, so it is necessary to convert the string vectors according to the dictionary, and place each word as an element of the converted list. Each element in this list has also a grammatical class, so the result is a list of tuples. The first element is a number representing a word and the second represents a POS-Tag. The POS-Tag number will be the list vector, where each list represents a text and each element represents a word with its POS-Tag. This input set will be the independent variable in the artificial neural network.

To create the dependent variable, you must convert the entities of each token to numbers. The new list of numeric tokens will be related to the "tag2idx" dictionary. Each entity in this list is related to an entity in the independent variable vector list.

When the token has no localized entity, it is classified as "O". If the token has a localized entity, it is classified with a letter according to its position. If the token is the first token of the entity, it will have the letter "B" (beggin) + "−" + "entity tag". If the token is not the first one, it will have the letter "I" + "−" + "entity tag". The entities tokens defined are:

1. B-LOC: first local entity token, i.e.: "São"
2. I-LOC: remaining location tokens, i.e.: "Paulo"
3. B-ORG: first entity token Organization, i.e.: "Banco"

4. I-ORG: remaining organization tokens, i.e.: “Votorantim”
5. B-PER: first person entity token, i.e.: “Paulo”
6. I-PER: remaining person token, i.e.: “de”, “Tarso”
7. O: unclassifiable remaining tokens, i.e.: “Transferência”

Finally, we need to standardize the size of the lists present in the vector, since the artificial neural network needs all vectors to be the same size. For the standardization of vectors, it was found that the maximum size of tokens existing in the texts was of 75 characters. Thus, tokens smaller than 75 characters have been completed with zeros.

3.3 Training

The training module is composed by the parameters responsible for the creation of the artificial neural network such as number of layers, iterations, activation function, and loss function.

The artificial neural network used is called the recurrent neural network (RNN), which is composed of neural units called LSTM (long short-term memory). The LSTM neural unit is a variation of traditional artificial networks. It has the ability to persist information for an arbitrary time. This type of neural unit is widely used in problems where the existence of patterns has a long-term correlation. With RNNs it is possible to connect information from previous memory states to the current one. There are several applications for this type of artificial neural network, such as speech recognition, translation, image captioning, and natural language processing.

In our proposal, we use a LSTM neural unit specification, the bidirectional long short-term memory (BLSTM). As can be seen in Fig. 2, an RNN with BLSTM neural units connects the two-way neural network. Thus, prediction can be performed in both ways, with the aim of maximizing entropy.

Each LSTM unit consists of a cell, an input port, an output port, and a forgetting port that regulate the flow of information. The input port controls new information that can enter in the memory. The forgetting port controls when information should be forgotten by memory, allowing the cell to remember new data. The output port controls the information that leaves the cell. All cells are connected in bidirectional way, by performing forward and backward propagation in each way.

The training artifacts are: i) the binary file model to be reused without the need for retraining; ii) an output metric report against the validation base; iii) the execution time; iv) the parameters used to construct the model; v) the identification code of each text segment of the database, which were used during the process; vi) the history of each iteration of the network along with the loss metrics; vii) the error; viii) a file containing all tokens used in the validation, the true value, and the ones sorted by model.

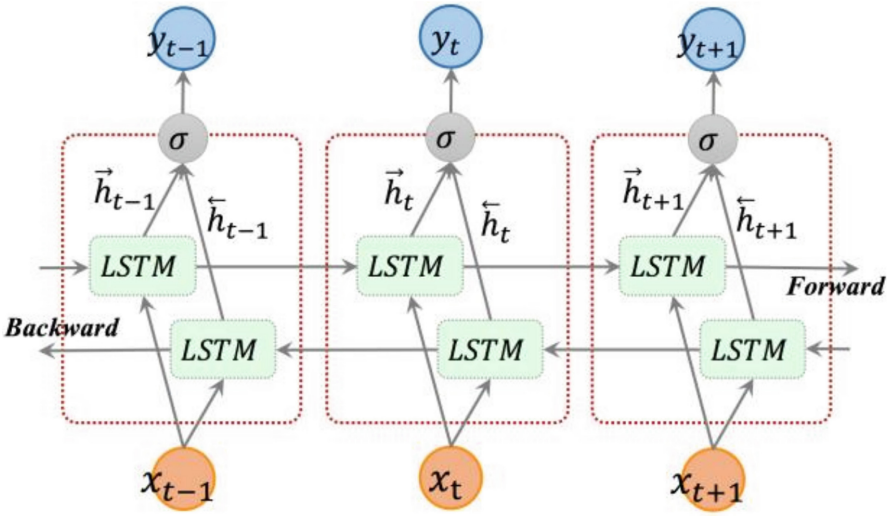

Fig. 2. Recurrent Neural Network with bidirectional long short-term memory

3.4 Version Control

The version control module is responsible for loading the saved data, which includes: the model, the word dictionary, the character dictionary, the tag dictionary, data converted to number, and the conversion to text format again. The main function of this module is to maintain a version history of models created from the training process. Therefore, it is possible to verify the evolution of a model in relation to the accuracy, and thus identify the changes that were significant. Therefore, it is possible to verify the evolution of the model, in relation to the accuracy, and identify the changes that were significant.

4 Methodology Assessments

In this section, we present the assessments of the proposed methodology. The amount of text snippets varied for each entity token, as well as the number of epochs of the training algorithm. As is shown in Fig. 3.

The first evaluation was run with 18001 text snippets and 40 training epochs. The general amount of text snippets was increased and the training epochs was decreased. The main goal was to verify the influence of the size of dataset/corpus for training and the number of epochs required to perform in order to recognize entities.

4.1 Results

In our earlier results, as can be seen in Fig. 4, the accuracy increased for most tokens, from test 1 to test 4:

Test	Tranning Tracks	Epochs	Execution Time (minutes)	Accuracy	Accuracy Isolated
1	18001	40	1334	98,52%	72,65%
2	19001	20	43	98,69%	76,28%
3	28001	20	74	97,23%	83,65%
4	50001	20	147	97,54%	85,64%

Fig. 3. Assessments sets

- B-LOC from 83,44% to 93,36%
- B-ORG from 82,88% to 82,62%
- B-PER from 82,20% to 88,50%
- I-ORG from 25% to 86,61%
- I-PER from 26,53% to 90,91%
- I-LOC from 84,62% to 87,50%

The recognition for I-PER had the highest increased of accuracy around 64%, followed by I-ORG, which the accuracy was increased around 60%. The accuracy for B-LOC was increased around 10%, and B-PER and I-LOC accuracies were increased by around 6% and 3%, respectively. The exception was for recognizing B-ORG entities, to which the accuracy has remained largely the same, with a slight decrease of less than 1%. Regarding the corpus chosen for the tests, the I-LOC entity token was not presented nor in test 1, neither in test 2.

Moreover, the O entity token had also largely the same accuracy during the tests. This kind of entity means represents unclassified token, i.e., in the end, the proposed methodology possess a high accuracy in differentiating words that are not named entities.

In addition, we calculated the isolated accuracy, which means overall correct recognition for all entities, excluding entity O. The isolated accuracy was increased from 72,65% to 85,64%, as the amount snippets was increased, and the total accuracy was kept nearly the same, as it can be seen in Fig. 5. Furthermore, by decreasing the epochs in the training algorithm, the execution time also decreased, as shown in Fig. 3. Most of cloud services capable of processing this kind of training are paid, consequently, the reduction of time and the consumption of computational resources are relevant points to be considered.

5 Final Remarks and Future Works

The preliminary results are promising. It has been proved that regardless of the language domain of the corpus, text parsing is possible. By means of the chosen machine learning algorithm, it was possible to create models that have assertiveness relatively close to those currently obtained by English language-based models.

The biggest challenge is in relation to the annotations, because during the research we found several corpora in Portuguese, mostly incorrect, incomplete or, when complete, composed of insufficient expressive data.

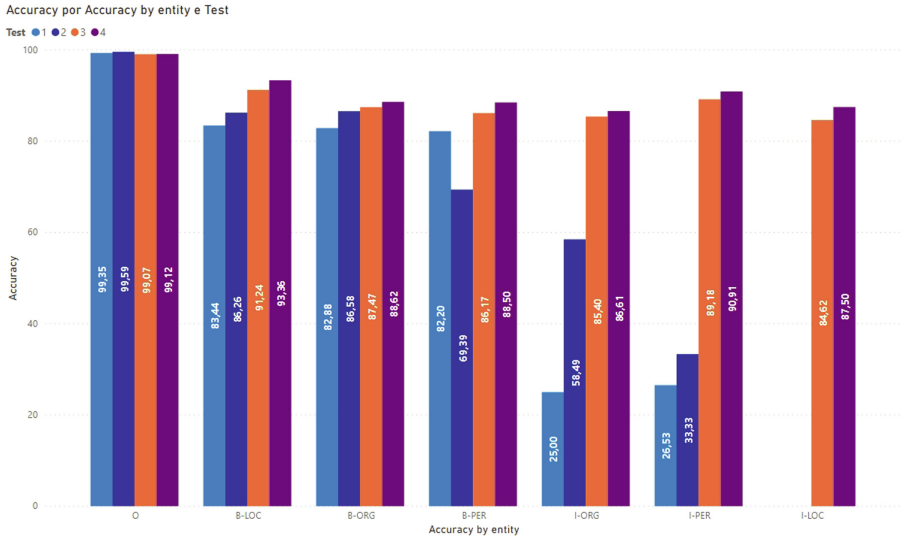

Fig. 4. Accuracy by entity token

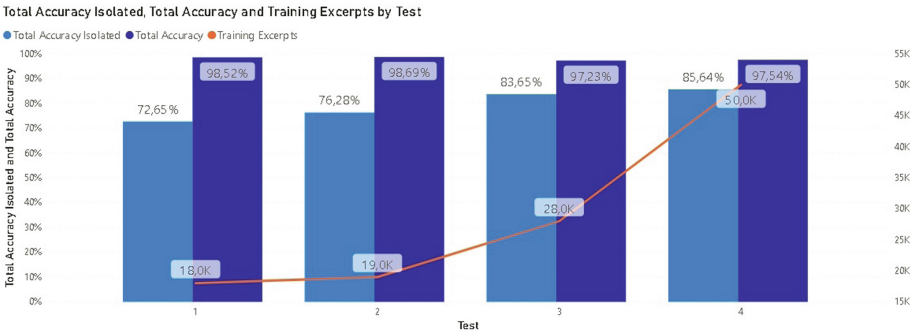

Fig. 5. Total and isolated accuracy

Despite the challenges encountered during this work, it can be said that the overall goal was achieved, since it was possible to create a methodology that was able to extract entities with results close to other results in other languages, such as English, which possess less morphological and syntactic complexity.

The next steps for this research will firstly increase the number of snippets to cover all entities during all test iterations. In the end, we intend to build a unique corpus by assembling the Portuguese corpora available.

References

1. Abdelali, A., Cowie, J., Soliman, H.: Building a modern standard Arabic corpus. In: Workshop on Computational Modeling of Lexical Acquisition, pp. 25–28 (2005)

2. do Amaral, D.O.F., Vieira, R.: NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. Linguamática **6**(1), 41–49 (2014)
3. Amri, S., Zenkouar, L., Outahajala, M.: Build a morphosyntaxically annotated amazigh corpus. In: Proceedings of the 2nd International Conference on Big Data, Cloud and Applications, p. 8. ACM (2017)
4. Cavalin, P., et al.: Building a question-answering corpus using social media and news articles. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 353–358. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_36
5. da Consolação Dias, C.: A análise de domínio, as comunidadesdiscursivas, a garantia de literatura e outras garantias. Informação Sociedade **25**(2) (2015)
6. Holthaus, P., et al.: How to address smart homes with a social robot? A multimodal corpus of user interactions with an intelligent environment. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 3440–3446 (2016)
7. IBGE: Acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal (2017). https://biblioteca.ibge.gov.br/index2.php/biblioteca-catalogo?view=detalhes&id=2101631
8. Mooers, C.N.: The next twenty years in information retrieval; some goals and predictions. Am. Doc. **11**(3), 229–236 (1960)
9. Mosley, M., Brackett, M.H., Earley, S., Henderson, D.: DAMA Guide to the Data Management Body of Knowledge. Technics Publications (2010)
10. Nelson, D.M.Q.: Uso de redes neurais recorrentes para previsão de séries temporais financeiras (2017)
11. Olah, C.: Understanding LSTM Networks (2015)
12. de Oliveira, M.G., de Souza Baptista, C., Campelo, C.E., Bertolotto, M.: A gold-standard social media corpus for urban issues. In: Proceedings of the Symposium on Applied Computing, pp. 1011–1016. ACM (2017)
13. Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R.R., Stauffer, M., Couto, S., Bermejo, P.: LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In: Villavicencio, A., et al. (eds.) PROPOR 2018. LNCS (LNAI), vol. 11122, pp. 313–323. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99722-3_32. https://cic.unb.br/~teodecampos/LeNER-Br/
14. Pirovani, J., Oliveira, E.: Portuguese named entity recognition using conditional random fields and local grammars. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
15. Pretto, J.R.: O estilo jornalístico. Estudos Linguísticos **38**(3), 481–491 (2009)
16. Santos, J., Consoli, B., Santos, C., Terra, J., Collonini, S., Vieira, R.: Assessing the impact of contextual embeddings for Portuguese named entity recognition, pp. 437–442, October 2019. https://doi.org/10.1109/BRACIS.2019.00083
17. Souza, F., Nogueira, R., Lotufo, R.: Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:1909.10649 (2019)
18. Spoustová, J., Spousta, M., Pecina, P.: Building a Web Corpus of Czech (2010)
19. Villalva, A., Mateus, M.H.M.: Morfologia do português. Universidade Aberta Lisboa (2008)
20. Zhou, P., et al.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 207–212 (2016)