Adapting NER (CRF+LG) for Many Textual Genres *

Juliana Pirovani[1], James Alves[2], Marcos Spalenza[2], Wesley Silva[2], Cristiano da Silveira Colombo[2], and Elias Oliveira[2]

[1] Universidade Federal do Espírito Santo (UFES), 29.500-000 - Alegre - ES - Brasil
`juliana.campos@ufes.br`
[2] Programa de Pós-Graduação em Informática Universidade Federal do Espírito Santo (UFES), 29.075-910 - Vitória - ES - Brasil
`{james,elias}@lcad.inf.ufes.br`

Abstract. Named Entity Recognition is the task of automatically identifying named entities and classifying them into predefined categories such as *person, place, organization*, among other categories considered relevant in specific domains. This task is important and challenging, especially when the system must be able to recognize named entities in many textual genres, including genres that differ from those for which it was trained. CRF+LG is a hybrid system for Named Entity Recognition in Portuguese texts that combines a labeling obtained by a Conditional Random Fields with a term classification obtained by a Local Grammar as an additional informed feature. This paper aims to report the initial efforts made to adapt CRF+LG system for many textual genres in accordance with the proposed Portuguese Named Entity Recognition task in IberLEF 2019. We adapted the LG to capture rules of textual genres that do not appear in the examples of the training corpus and thus assist the Named Entity Recognition, even when there is no training set of an available textual genre. CRF+LG was also trained in an augmented training corpus.

Keywords: Named Entity Recognition · Conditional Random Fields · Local Grammars · Domain Adaptation

1 Introduction

Named Entity Recognition (NER) is a task for identifying and classifying automatically named entities (NEs) in free written texts. These NEs correspond to names of *person, places, organizations*, among other categories considered relevant in specific domains. This task is important because it is a fundamental step

of preprocessing for several applications such as question answering systems [20], relation and event extraction [5] and entity-oriented search [6]. Indeed, NEs are an essential source of information in textual information retrieval.

NER is a very challenging task as several categories of named entities are written similarly and they appear in similar contexts. In addition, NER depends on the language, the training corpus and a given domain [17]. Considering the domain dependency, the same category of NE can be written in different ways depending on the textual genre under analysis. For example, in e-mail texts it is common to see *person* names after words as *Hello* and *Good afternoon*, whereas in memorandum texts it is common to see *person* names after words as *Public servants* and *Professor*. Consistent training sets including texts from different genres are not always available.

In 1995, the Message Understanding Conference [13] included the NER task for the first time for the English, carrying out a joint assessment of the area. Thereafter, several similar events have emerged such as the ACE [8], CoNLL [24], HAREM [12, 27] and TAC [14]. HAREM was an initiative for the Portuguese organized by Linguateca [11]. The annotated corpora used in the First and Second HAREM, known as the Golden Collections (GC), are used as a golden standard reference for NER systems in Portuguese.

This year (2019), the Portuguese NER task was one of the tasks proposed in the Iberian Languages Evaluation Forum (IberLEF) [23]. The objective of this task is to evaluate the submitted systems in many textual genres. The participants were free to choose their own training datasets. The categories *person, place, organization, value* and *time* were evaluated in datasets that have as main textual genres: news, memorandums, e-mails, interviews and magazine articles; and the *person* category was evaluated in clinical notes and police texts.

This paper presents the initial efforts to adapt the system CRF+LG [21] for many textual genres in accordance with this proposed task in IberLEF 2019. CRF+LG is a hybrid system for Portuguese NER that combines a labeling obtained by a Conditional Random Fields (CRF) with a term classification obtained by a Local Grammar (LG) as an additional informed feature. The idea of this system was to study a way to improve the performance of NER systems that use the machine learning approach using less training corpus. In order to participate in the IberLEF 2019, we observed some datasets from different textual genres, we also adapted the LG and retrained the model with an augmented training corpus.

The remaining of this paper is organized as follows. In Section 2 we discuss some of the more related works which both support some of our arguments and complement some point of view we discuss in this paper. The methodology is explained in the Section 3. Within this section we enumerate each of the necessary steps to perform the training and testing and we describe the adaptations made in this architecture to the IberLEF. We also introduce some challenges we had found within the datasets used for training which decrease the performance of the learning process. The Section 4 discusses the results yielded by our algorithm which was run by the IberLEF organizers. We also discuss some aspects faced

when dealing with cross domain datasets. Our conclusions are presented in the Section 5.

2 Related work

Named Entity Recognition systems can be developed using the following approaches: linguistics [17, 22], machine learning [4, 25, 29] or hybrid [19, 30]. Some of the main NER systems for Portuguese will be described below.

The system proposed by [25] is based on the CharWNN Deep Neural Network, which uses word-level and character-level representations to perform sequential classification. The system was tested for the Portuguese and Spanish and, for the Portuguese, the GC of the First HAREM was used as training set and the MiniHAREM as the test set. The approach was compared to the ETL_{CMT} system [26], an ensemble method based on Entropy Guided Transformation Learning (ETL) and outperformed this system in both total (10 categories of HAREM) and selective (categories *person, place, organization, time* and *value*) scenarios.

A Deep Neural Network architecture with word-level and character-level representations was also used in [4]. A combination of these representations is fed into a bidirectional Long Short-Term Memory with Conditional Random Fields (Bi-LSTM-CRF) to perform sequential classification. The authors evaluated different combinations of hyperparameters for training such as word embeddings model, tagging schemes, word capitalization feature and number of hidden units for each LSTM, obtaining the optimal values for the parameters that had a greatest impact in the performance of the model. A very similar architecture was used by [7] for two sequence labeling tasks (POS-tagging and NER) obtaining very close results.

A hybrid approach to Portuguese NER is presented in [18, 21] using the machine learning approach CRF [10] and the linguistics approach LG [9]. The classification obtained from LG was sent as an additional feature for the learning process of the CRF prediction model. The CRF model assigns the final label of the NEs. This approach is a good way to take into account the human expertise for capturing the rules that do not appear in examples of the annotated corpus used for training by the CRF. A study about the boundaries of CRF's performance when using a result coming from any other classifier as an additional feature was also presented.

The systems that used Neural Networks [4, 7, 25] presented superior results using massive corpora for unsupervised learning of features, which was not the case of the work presented in [21]. However, the results obtained by [21] outperform the results of systems reported in the literature that were evaluated under equivalent conditions: a system that uses only CRF [1] and the system based on the CharWNN presented in [25] without the unsupervised pre-training.

3   Methodology

In order to participate in the IberLEF 2019, we have used the architecture of our system CRF+LG[21]. CRF+LG does not use massive corpora for unsupervised learning of features. The LG is a good way to take into account the human expertise for capturing the rules and a way to perform the NER using the linguistic approach when there is no available training corpus. The Figure 1 presents an overview of the methodology used, demonstrating how the steps to perform the training occur.
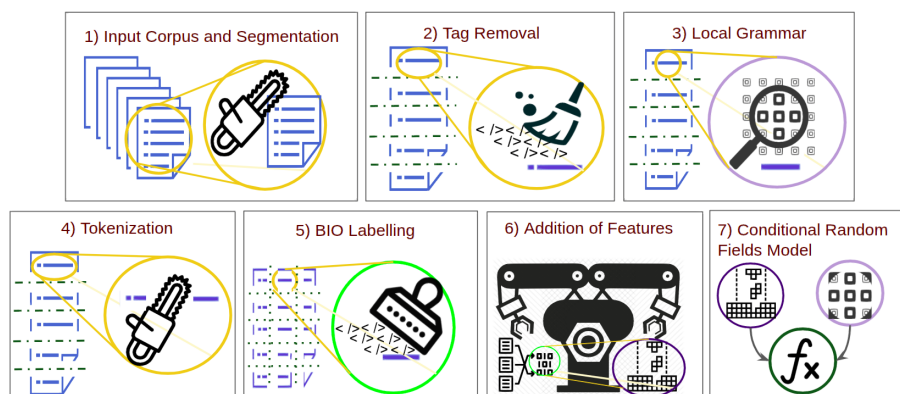

Fig. 1. Train Workflow

Initially, each input file goes through the sentence segmentation process (step 1). Segmentation was performed using the Unitex (http://unitexgramlab.org/) tool. Unitex uses LGs to describe the different ways that indicate the end of a sentence. For this work, the LG that performs sentence segmentation in Unitex has been changed so as not to segment sentences in a colon (:) and semicolon (;). This flexibility is a strength of the tool.

A copy of the targeted files has their tags removed since the CD used has the NEs markings (step 2). The LG built in this work is applied to these files without any marking and the NEs identified by it are annotated (step 3). On the other hand, the segmented files are tokenized using the OpenNLP (http://opennlp.apache.org/) library (step 4). This library is based on machine learning and performs common NLP tasks such as segmentation, tokenization, POS-Tagging, etc.

In order to represent the NER as a sequence labeling problem, a label must be assigned to each token of the text. The BIO notation was used (steps 4 and 5). In the sequence, several features [18] are added for each token of the files, including the NE label previously assigned by the LG (step 6). These characteristics are used during supervised learning of the CRF prediction model (step 7).

The methodology used for testing is similar, but the input files do not have the NEs tags. In addition to the files containing the tokens and features, the CRF receives the previously trained model to predict a label for each token.

The next two sections have a short description of how the system obtains a tip by the LG and explain how CRF works. In the last section we described the adaptations made to participate in the IberLEF event.

3.1 Local Grammars (LG)

An LG created in Unitex is represented as a set of one or more graphs. The LG used by CRF+LG consists of 10 graphs, one for each of the NEs categories considered by HAREM.

We observed in the training file in which context each type of NE appeared, what words could somehow indicate the existence of NE to construct each graph. We observed that, for example, words with the first letter capitalized preceded by the preposition *em* (in) were labeled as *place*. We also observed that some NEs of the *person* category are preceded by words such as *diz* (say), *explicou* (explained), *afirmou* (said), etc.

Thus, the graphs created capture some simple heuristics to the recognition of NEs in the training set. An example of rule in the graph created for the *person* category is presented in Figure 2.
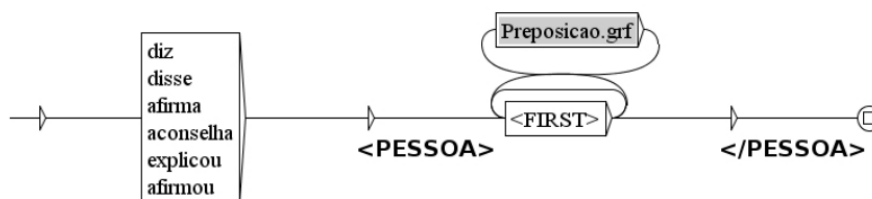

Fig. 2. Example of rule in the graph that recognizes the Person category

This graph recognizes words such as *diz* (say) or *afirmou* (said) followed by words with the first letter capitalized, as identified by the code $<$ FIRST $>$ in Unitex dictionaries. Among words with the first letter capitalized, prepositions may appear whose recognition has been previously detailed in graph Preposicao.grf included as subgraph. Examples of occurrences identified by this graph were: diz $<$ PESSOA $>$ Moncef Kaabi $<$ /PESSOA $>$
afirmou $<$ PESSOA $>$ José SÓCRATES $<$ /PESSOA $>$
afirma $<$ PESSOA $>$ Jason Knight $<$ /PESSOA $>$.

Note that identified person will appear between the tags $<$ PESSOA $>$ ($<$ PERSON $>$) and $<$ /PESSOA $>$ in the concordance file containing the list of occurrences identified.

3.2 Conditional Random Fields (CRF)

Conditional Random Fields (CRF) is a machine learning method for structured prediction proposed by [10]. It is used for labeling of sequential data based on a conditional approach.

Let $X = (x_1, x_2, ..., x_n)$ be a sequence of words in a text, we want to determine the best sequence of labels $Y = (y_1, y_2, ..., y_n)$ for these words, corresponding to the categories of NEs (10 categories of the HAREM or the label O in this work). The CRF models a conditional distribution $p(Y|X)$ that represents the probability of obtaining the output Y given the input X.

In this work, we used a linear-chain CRF that predict the output variables Y as a sequence for sequences of input variables X. According to [28], a linear-chain CRF is a conditional distribution that takes the form shown in Equation 1:

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp \left\{ \sum_{k=1}^{K} \theta_k f_k \left(y_t, y_{t-1}, \mathbf{x}_t \right) \right\} \tag{1}$$

where $Z(x)$ is a normalization function given by Equation 2

$$Z(x) = \sum_{y} \prod_{t=1}^{T} \exp \left\{ \sum_{k=1}^{K} \theta_k f_k \left(y_t, y_{t-1}, \mathbf{x}_t \right) \right\} \tag{2}$$

$F = \{f_k(y_t, y_{t-1}, x_t)\}_{k=1}^{K}$ is a set of feature functions that must be fixed according to the problem. An example is a function which takes the value 1 when the word begins with a capitalized letter (component of the input vector x_t), its label is *Person* (y_t) and the previous label (y_{t-1}) is Other and 0 otherwise. The vector x_t contains all the components of the global observations x that are needed for computing features at time t. $\theta = \theta_k$ is a vector of weights that must be estimated from the training set. This is usually done by maximum likelihood learning. The weights depend on each feature function and the more discriminating the function, the higher its computed weight will be.

The MALLET (http://mallet.cs.umass.edu/) toolkit was used in this work to estimate the vector of weights and then apply the CRF model obtained to label the test set. This CRF model combines the weights of each feature function to determine the probability of a certain value (y_t).

3.3 Adaptation of CRF+LG to IberLEF

CRF+LG was built to recognize the 10 named entities categories of the HAREM (*person, place, organization, value, time, event, abstraction, work, thing* and *other*). Then, the system was initially adapted to consider only the five categories of the IberLEF (*person, place, organization, value and time*) during the CRF training phase. Nevertheless, we have kept the recognition of the 10 categories by the LG because we believe that this helps the system to disambiguate NEs.

The Golden Collection of the First and Second HAREM, considered as a reference for Named Entity Recognition systems in Portuguese, were used in previous experiments [19] as training and testing sets, respectively, for evaluation of the CRF+LG. Several errors occurred due to some inconsistencies in the GC of the First HAREM and Second HAREM. For example, in the GC of the First HAREM, strings as 2004 preceded by the preposition *em* (in) are considered NEs of the Time category and the CRF+LG learned this and labeled all similar strings preceded by *em* as Time. However, in the GC of the Second HAREM, the preposition *em* is part of the NE. So all these NEs were wrongly labeled. The same happened in other situations of the categories *time, value* and *person*.

Some of these major inconsistencies were removed by Pirovani [21] and others were removed during this work. The goal was to get a more consistent dataset, normalized, composed of the three GCs of the HAREM (First HAREM, Mini HAREM and Second HAREM) to use as training.

The GCs of the HAREM include documents from different textual genres such as news, web texts, literary fiction, transcribed oral interviews, technical texts, journalistic and personal blog, essays and FAQ questions [12, 27]. However, the task of the IberLEF proposes to evaluate the systems in other specific textual genres such as memorandums, e-mails, magazine articles, clinical notes and police texts.

In order to train CRF+LG to this task, we have researched and reviewed other corpus from different textual genres in Portuguese:

1. SIGARRA [16]: SIGARRA corpus has 905 articles, manually annotated using eight NEs categories: *hour, event, organization, course, person, location, date* and *organic unit.*
2. WikiNER [15]: This corpus is a silver-standard automatically annotated containing three different NEs annotated: *person, location* and *organization*. We created 592 subsets and reviewed 40 parts including annotation for value and time for NEs and adjusting the automatic annotation mistakes.
3. LeNER-BR [2]: LeNER-BR was manually annotated with a focus on legal documents. This dataset has 70 documents with the following categories of NEs: *organization, person, time, locations, law* and *decisions* regarding law cases.
4. aTribuna [21]: This dataset has 100 newspaper documents with 2714 NEs *person* manually annotated.
5. administrative orders (http://gedoc.ifes.edu.br/): We also annotated manually 20 administrative orders of the Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo (IFES).

Our initial intention was to use these datasets to 1) identify new rules to insert into LG and 2) combine them to increase the training set and thus improve the model prediction. However, some inconsistencies observed between the GCs of the HAREM and others such as LeNER and SIGARRA made it difficult to integrate all these datasets to create a unique training set.

The LG used in CRF+LG was built by analyzing only the CD of the First HAREM. By analyzing some texts of these new domains, we observed some very

strict patterns for writing of NEs and several adaptations have been introduced at LG to recognize these patterns. Here are some examples:

1. Sequences of words with the first letter capitalized or numbers beginning with words such as *Sala, Salão, Auditório* and *Anfiteatro* as *place* category.
2. Recognition of dates (*time* category) with dots (25.12.2010).
3. Recognition of dates preceded by words such as *até, a partir de, entre, dia* and *desde.*
4. Recognition of values preceded by abbreviations or words such as *num. N., art. Art., matrícula* and *siape.*

One of the main inconsistencies observed among the datasets was the different categories of NEs annotated. For example, the SIGARRA corpus does not contain the *value* category annotated, however there are NEs of this category in the texts. Another example of inconsistency are the NEs annotated in different ways. Sometimes specific words in lowercase letters should form part of NEs and other times not. For example, *rainha* (queen) in *rainha Elizabeth* (queen Elizabeth) and *mais de* (more than) in *mais de 30* (more than 30). This certainly deteriorate the model learning because of the lack of correct or consistent annotation.

4 Experiment Result

Before submitting the system to the IberLEF, we repeated some of the experiments performed in [21]. Initially, the LG built in [21] and the new version of our LG submitted to IberLEF were applied individually to the GC of the Second HAREM to evaluate the new rules inserted.

Although the precision value obtained by adapted LG was lower indicating that more NEs have been misidentified (false positives) due to the new rules, these rules also increased the recall value in 9 percentage points. Thus, the gain obtained by adapted LG in comparison to the original LG was approximately 7 percentage points in F-measure. The decrease in the precision metric is some of the effect faced when we change the domain of the dataset used for testing. This experiment only suggests that the continuing adaption of the LG is a necessity.

CRF+LG was also rerun using the adapted LG. The GCs of the First HAREM and Second HAREM were used as training and testing sets respectively. The final gain in F-measure was about 4 percentage points achieving 63.11% in F-measure. These results are another example of how the combination CRF+LG can improve the NER. In this experiment we were able to identify 31 more entities due to the new version of the LG.

We also performed some experiments combining several of the datasets presented in the previous Section (GCs HAREM normalized, SIGARRA, selected sentences from WikiNER, aTribuna and administrative orders) for use as training set. The CRF+LG prediction models were obtained for all combinations and applied in a testing set that we have created for this purpose. This dataset contains only 15 texts from different textual genres annotated. The model that

presented the best results in this initial test was submitted to IberLEF. This model was trained with the GCs HAREM normalized and the 20 administrative orders.

4.1   IberLEF Task Results

The IberLEF organizers evaluated the submitted systems in two manually annotated datasets: the *Clinical dataset* with 50 sentences and 77 NEs and the *Police dataset* from the Brazil's Federal Police with 1388 sentences and 916 NEs. Both datasets were annotated with only the *person* category. The systems were also evaluated in the *General dataset* containing the SIGARRA dataset with NEs categories *date* and *time* mapped to as a single category *time* and a subset of sentences from the GC of the Second HAREM (SecHAREM) annotated with only the *value* category since SIGARRA does not have this category annotated.

The IberLEF organizers used the precision (P), recall (R) and F-measure (F) [3] metrics and computed the results using the CoNLL-2002's standard evaluation script (http://www.cnts.ua.ac.be/conll2002/ner/bin/conlleval.txt). The results to our model are exposed in Table 1.

Corpus	Category	P	R	F
Police Dataset	PER	29.59%	58.41%	39.28%
Clinical Dataset	PER	14.29%	10.09%	11.83%
General Dataset (SIGARRA + SecHAREM)	Overall	56.26%	56.66%	56.46%
	ORG	42.27%	32.31%	36.63%
	PER	57.39%	62.14%	59.67%
	PLC	37.35%	51.38%	43.26%
	TME	71.33%	74.91%	73.08%
	VAL	80.19%	82.52%	81.34%

Table 1. IberLEF Task Results

In the first column, we have the list of datasets: the *Police dataset* in the first line, followed by *Clinical dataset*, and the combined *SIGARRA +SecHAREM*. Whereas for the two first datasets only the *person* entity was evaluated, for the combined dataset all the five entities were evaluated: *ORG – organization*, *PER – person*, *PLC – location*, *TME – time* and *VAL – value*.

The best result obtained by our approach was on the identification of the *value* category (81.34% in F-measure) in the last line of the Table 1 for the *General dataset*, whereas our worst result was on identifying the *person* category for the *Clinical dataset*, in the second line (11.83% in F-measure).

Note that, based on the results depicted in Table 1, our approach did not achieve the same figures level on the two first datasets as we could get on the *Overall* evaluation when testing on the combined dataset. Although these datasets (*Police* and *Clinical*) were not divulged by the IberLEF organization

because the information is of a sensitive nature, we imagine that these results are due to NEs with structures very different from those for which this system was trained. The *Clinical dataset*, for example, has a textual structure with words that should be separated by a space and they are not, several medical abbreviations of unusual terms and odd sequences of special characters (AnaR1 and ###Paulo as *person* names). In order to recognize these very specific structures the system would need to be trained in texts from that same domain or have knowledge of those structures to insert into LG.

The results obtained in the *General dataset* were a bit better. The results for the *value* category exceeded 81 percentage points in F-measure and for the *time* category exceeded 73 percentage points in the same metric. NEs of these categories have better defined structures that are easier to capture in the LG rules and easier to learn by the CRF.

In order to understand our results better, we applied the CRF+LG model to the *General dataset* (https://github.com/jneto04/iberlef-2019) released by the organization. By analyzing the results obtained, we observed that many of the NEs of the *value* category have words such as *mais de* (more than), *cerca de* (about), *aproximadamente* (approximately) and *until* (até) which should be part of the NE. However, with the purpose of normalizing the three GCs of the HAREM to use as a training set, these words were removed. So, instead of recognizing sequences such as *mais de 800 milhões, cerca de 600 km, aproximadamente 1,4 tonelada e até 120 kg*, CRF+LG recognized *800 milhões, 600 km, 1,4 tonelada e 120 kg*, decreasing the value of the metrics.

CRF+LG recognized sequences preceded by words such as *Faculdade* (College), *Universidade* (University), *Instituto* (Institute) and *Departamento* (Department) as *organization* (Faculdade de Ciências Médicas da Universidade Nova de Lisboa, Departamento de Química). However, the IberLEF organization did not consider the *organic unit* category of the SIGARRA as an *organization*.

We also believe that the use of the 20 administrative orders as training set may have somewhat impaired the recognition of words in capital letters since many NEs are written in uppercase in this dataset.

It is important to note that the results obtained by the systems should not be directly compared as the participants used different training corpora. In this case, the CRF+LG also did not use massive corpora for unsupervised learning of features. In order to compare the techniques used by the systems, they must be trained in the same dataset and under equivalent conditions.

5 Conclusion

This paper is a result of the IberLEF task force which the objective is to evaluate intelligent algorithm models on the NER problem in many textual genres. Our proposed model used the combination of two strategies: a supervised learning algorithm named CRF, and a tailored set of LGs used here to give tips to the former algorithm. In [21] we discussed that the more valuable tips we offer to the CRF algorithm, the better is its performance.

In this paper we present the results yielded by the IberLEF organizers when running our model over the three datasets used to compare the participating systems. Two of these datasets were used solely to compute the performance of the submitted algorithms on automatically annotating the *person* entity on *police texts* and *clinical notes*.

The LG adapted in this work for use with the CRF+LG approach obtained a gain of 7 percentage points in F-measure in comparison to the original LG and a final gain of approximately 4 percentage points combined with the CRF according to the experiments presented in Section 4. These results show the potential of LG for use in the NER task and the necessity of the continuous adaptation of the LG.

The results obtained by the CRF+LG in the IberLEF task, especially for the *Police* and *Clinical datasets*, show the difficulty of the NER in new textual genres containing very specific structures that differ from those for which the system was trained. Our F-measure metric was below 12 percentage points in the *Clinical dataset* that presents particular challenges.

We observed some errors when analyzing the result obtained by the CRF+LG in the *General dataset* that could be avoided if we knew in advance which words should or should not be part of the NEs. In this way, LG and the training dataset could be tailored for this.

We claim that the IberLEF is a milestone towards on building a more uniform and better way to compare different approaches, measure their results and build better datasets for experimentation.

As a possible future work we think of better understanding how to decrease the impact of increasingly learning from a different domain. The idea is that a learning model from one domain can be cheaply used into another domain without a great impact observed in this paper. Besides, the preprocessing stage of the algorithms has also a great impact on the results. We are working on a way to introduce an intelligence layer within this stage in order to quickly learn the different textual genres and thus reduce the mistakes we also could find during the experiments carried out in this work.

References

1. Amaral, D.O.F.: O Reconhecimento de Entidades Nomeadas por Meio de Conditional Random Fields para a Língua Portuguesa. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brasil (2013)
2. Araujo, P., Campos, T., Oliveira, R., Stauffer, M., Couto, S., Bermejo, P.: LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text. In: International Conference on the Computational Processing of Portuguese (PROPOR). pp. 313–323. Lecture Notes on Computer Science (LNCS), Springer, Canela, RS, Brazil (September 24-26 2018)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca. Bookman Editora (2013)
4. Castro, P.V.Q., da Silva, N.F.F., da Silva Soares, A.: Portuguese Named Entity Recognition Using LSTM-CRF. In: Villavicencio A. et al. (eds) Computational

Processing of the Portuguese Language. PROPOR 2018. Lecture Notes in Computer Science, vol 11122. pp. 83–92. Springer, Cham, Canela, RS (Sep 2018)

5. Chan, Y.S., Roth, D.: Exploiting Syntactico-Semantic Structures for Relation Extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 551–560. Association for Computational Linguistics (2011)

6. Cheng, T., Yan, X., Chang, K.C.C.: Supporting Entity Search: a Large-scale Prototype Search Engine. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. pp. 1144–1146. ACM (2007)

7. Costa, P., Paetzold, G.H.: Effective Sequence Labeling with Hybrid Neural-CRF Models. In: Villavicencio A. et al. (eds) Computational Processing of the Portuguese Language. PROPOR 2018. Lecture Notes in Computer Science, vol 11122. pp. 490–498. Springer, Cham, Canela, RS (Sep 2018)

8. Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S., Weischedel, R.M.: The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In: LREC. vol. 2, p. 1. European Language Resources Association (ELRA), Lisboa, PORTUGAL (2004)

9. Gross, M.: The Construction of Local Grammars. In ROCHE, E.; SCHABÈS, Y. (eds.). Finite-state language processing, Language, Speech, and Communication, Cambridge, Mass. pp. 329–354 (1997)

10. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001. vol. 1, pp. 282–289. ACM, San Francisco, CA, USA (2001)

11. Linguateca: (2018), http://www.linguateca.pt/HAREM/, acesso em: 02/03/2018

12. Mota, C., Santos, D.: Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM. Linguateca (2008), https://www.linguateca.pt/LivroSegundoHAREM/

13. MUC-7: MUC-7 Proceedings (2016), acesso em: 11/10/2018

14. NIST: Text Analysis Conference (TAC) (2018), https://tac.nist.gov/2018/index.html, acesso em: 24/05/2018

15. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning Multilingual Named Entity Recognition from Wikipedia. Artificial Intelligence **194**, 151–175 (2013)

16. Pires, A.R.O.: Named Entity Extraction from Portuguese Web Text. Ph.D. thesis (2017)

17. Pirovani, J.P.C., Oliveira, E.: Extração de Nomes de Pessoas em Textos em Português: uma Abordagem Usando Gramáticas Locais. In: Computer on the Beach 2015. pp. 1–10. SBC, Florianópolis, SC (March 2015)

18. Pirovani, J.P.C., Oliveira, E.: CRF+LG: A Hybrid Approach for the Portuguese Named Entity Recognition. In: Abraham A., Muhuri P., Muda A., Gandhi N. (eds) Intelligent Systems Design and Applications (ISDA 2017). Advances in Intelligent Systems and Computing. vol. 736, pp. 102–113. Springer, Cham, Delhi, India (2017). https://doi.org/https://doi.org/10.1007/978-3-319-76348-4_11

19. Pirovani, J.P.C., Oliveira, E.: Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (may 2018)

20. Pirovani, J.P.C., Spalenza, M.A., Oliveira, E.: Geração Automática de Questões a Partir do Reconhecimento de Entidades Nomeadas em Textos Didáticos. In: XXVIII Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE 2017). vol. 28, pp. 1147–1156. Sociedade Brasileira de Computao - SBC, Recife, Brasil (2017)
21. Pirovani, J.P.C.: CRF+ LG: Uma Abordagem Híbrida para o Reconhecimento de Entidades Nomeadas em Português. Ph.D. thesis (2019)
22. Rocha, C., Jorge, A., Sionara, R., Brito, P., Pimenta, C., Rezende, S.: PAMPO: Using Pattern Matching and Pos-tagging for Effective Named Entities Recognition in Portuguese (2016), http://arxiv.org/abs/1612.09535
23. Sandra Collovini, Joaquim Santos, B.C.J.T.R.V.P.Q.M.S.D.B.C.R.G.C.C.a.X.: Portuguese Named Entity Recognition and Relation Extraction Tasks at IberLEF 2019 (2019)
24. Sang, E.F., 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 2003-Volume 4. pp. 142–147. Association for Computational Linguistics, Stroudsburg, PA, USA (2003)
25. Santos, C.N., Guimaraes, V.: Boosting Named Entity Recognition with Neural Character Embeddings. In: Proceedings of the Fifth Named Entities Workshop, ACL 2015. pp. 25–33. Association for Computational Linguistics, Stroudsburg, PA, USA (2015)
26. Santos, C.N., Milidiú, R.L.: Entropy Guided Transformation Learning: Algorithms and Applications. Springer-Verlag London, London, United Kingdom (2012)
27. Santos, D., Cardoso, N.: Reconhecimento de Entidades Mencionadas em Português: Documentação e Actas do HAREM, a Primeira Avaliação Conjunta na Área. Linguateca (2007), http://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf
28. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields. Foundations and Trends R in Machine Learning **4**, 267–373 (2012)
29. Yang, J., Zhang, Y., Dong, F.: Neural Reranking for Named Entity Recognition. arXiv preprint arXiv:1707.05127 (2017)
30. Zhang, B., Pan, X., Lin, Y., Zhang, T., Blissett, K., Kazemi, S., Whitehead, S., Huang, L., Ji, H.: RPI BLENDER TAC-KBP2017 13 Languages EDL System. In: Proceedings of the Tenth Text Analysis Conference (TAC2017). NIST, Maryland, USA (2017)