Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars

Juliana P. C. Pirovani, Elias de Oliveira
Universidade Federal do Espírito Santo - UFES
Av. Fernando Ferrari, 514, 29075-910 Vitória, ES, Brazil
juliana.campos@ufes.br,elias@lcad.inf.ufes.br

Abstract

Named Entity Recognition involves automatically identifying and classifying entities such as persons, places, and organizations, and it is a very important task in Information Extraction. Conditional Random Fields is a probabilistic method for structured prediction, which can be used in this task. This paper presents the use of Conditional Random Fields for Named Entity Recognition in Portuguese texts considering the term classification obtained by a Local Grammar as an additional informed feature. Local grammars are handmade rules to identify named entities within the text. The Golden Collection of the First and Second HAREM considered as a reference for Named Entity Recognition systems in Portuguese were used as training and test sets respectively. The results obtained outperform the results of competitive systems reported in the literature.

1. Introduction

Named Entity Recognition (NER) aims at automatically identifying and classifying entities such as persons, places, organizations and values. This is a fundamental task in Information Extraction since, besides having several applications, other tasks such as relations and events extraction, question answering systems and entity-oriented search depend on it as a preprocessing step (Jiang, 2012).

NER is not a simple task. Several categories of named entities (NEs) are written similarly and they appear in similar contexts. In addition, the same NE can be classified into different categories depending on the surrounding context and some entities do not appear even in large training sets. Thus, dictionaries are not always useful.

In 1995, the Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996) included the NER task for the first time for the English. Several similar events have emerged later such as the CoNLL (Tjong Kim Sang and De Meulder, 2003) and the HAREM (Santos and Cardoso, 2007; Mota and Santos, 2008). HAREM was an initiative for the Portuguese and 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.

HAREM differs from other similar events in two aspects (Mota and Santos, 2008): the classification of an NE depends only on its use in context and more than one classification can be assigned to an NE. Moreover, the HAREM classifies 10 categories of NEs (Person, Place, Organization, Value, Time, Event, Abstraction, Work, Thing and Other). Thus, the HAREM presents a more demanding task and, therefore, the performance values obtained using its reference data sets are still lower compared to the others (Santos and Cardoso, 2007).

NER systems can be developed using the following approaches: linguistics, machine learning or hybrid. This work seeks to explore the potential of the linguistics and machine learning approaches by constructing a hybrid system for NER in Portuguese. The presented strategy, CRF+LG, combines a labeling obtained by a Conditional Random Fields (CRF) with a term classification obtained from Local Grammars (LGs). Sutton and McCallum in (Sutton and McCallum, 2012) say that an interesting type of feature for the CRF can be the result of simpler methods for the same task. Thus, in this work we apply LGs to perform the pre-labeling by capturing general evidence of NEs in texts and the CRF performs sequential labeling using this pre-labeling. The pre-labeling is sent to the CRF together with other features and can be seen as a suggestion for the CRF.

Conditional Random Fields (Lafferty et al., 2001) is a machine learning method, which has been successfully used in several Natural Language Processing (NLP) tasks, including NER. NER is treated as a sequence labeling problem and a conditional model is constructed from a training data set to predict which is the best labeling sequence given an input sentence.

Local Grammars are one means of representing the contextual rules of the linguistics approach. ”Local grammars are finite-state grammars or finite-state automata that represent sets of utterances of a natural language” (Gross, 1999).

This paper is organized in 5 sections. Section 2 presents the state of the art and the Section 3 presents the methodology used in this work. The results of the study are presented in Section 4 and Section 5 presents conclusions and future works.

2. State of the Art

The systems presented in (Amaral, 2013) and (Santos and Guimaraes, 2015) achieved the best results for the 10 categories of the HAREM to date.

The NERP-CRF system, based on Conditional Random Fields (CRF), achieved the best Precision and F-Measure results compared to systems of the Second HAREM for the 10 categories (Amaral, 2013). NERP-CRF was also one of the four tools used to recognize NEs in Portuguese texts compared in (Amaral et al., 2014). The system obtained the best Precision results and the best performance for the Organization class considering only Person, Place and Organization categories.

In (Santos and Guimaraes, 2015), the authors proposed a language-independent system based on the CharWNN Deep Neural Network (DNN), which uses word-level and character-level representations to perform sequential classification. The approach presented better results compared to the ETL$_{CMT}$ system, an ensemble method that uses Entropy Guided Transformation Learning (ETL).

A combination of K-Nearest Neighbors (KNN) and CRF for English NER in tweets was proposed in (Liu et al., 2011). Due to insufficient information in a tweet and the unavailability of training data, a semi-supervised learning and 30 gazetteers were used. The KNN classifier conducts a word-level classification and the labeled results are fed into a CRF together with other conventional features. The KNN and CRF models are repeatedly retrained with an incrementally augmented training set. The method showed advantages over the baselines.

In (Constant and Tellier, 2012), the authors propose to evaluate the impact of external lexical resources into a CRF in order to perform the joint task of multiword segmentation and part-of-speech tagging in French. The information coming from dictionaries and local grammars recognizing numerical determiners and some NEs like organization or place was coupled as features into a CRF in two different ways: concatenating each possible POS category (Learn-concat) and considering each possible category in the resources as a new boolean property (Learn-bool). They obtained a gain of 0.5% in terms of F-measure and showed that the integration of lexicon-based features significantly compensates the use of a small training corpus.

This paper aims to perform the NER for the 10 categories of the HAREM using CRF as it was carried out in (Amaral, 2013); however, the preprocessing of the texts was performed differently and an initial information about the label of each word was obtained by LGs and added to the feature set sent to the CRF training phase.

This work also differs from those presented in (Liu et al., 2011) and (Constant and Tellier, 2012) by combining a rules-based approach with CRF for Portuguese NER and by not using gazetteers or dictionaries. To the best of our knowledege, there is not yet a work that combines LG (or other lexical resources) and CRF for NER in Portuguese.

3. The Methodology

In this work, the GC of the First and Second HAREM were used as training and test sets respectively. Both GC have 129 texts written in Portuguese and are available in (Linguateca, 2017).

During the training phase, initially each input file is splitted into sentences by the tool Unitex[1]. 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 was changed to not split the sentences in a colon (:) and a semicolon (;).

A copy of the segmented files has their tags removed since the GC used has the NEs tags. An LG is applied to these files without any markup and the NEs identified by it are annotated.

On the other hand, the segmented files are tokenized using the OpenNLP[2] library. In order to represent the NER as a sequence labeling problem, a label must be assigned to each text token. Several notations can be used to delimit NEs and identify tokens in text (Konkol and Konopík, 2015), but the IO notation was chosen because it presented better results in previous tests performed during this work.

The IO notation is used as follows: all tokens which are part of the NE are then labeled with I (Inside) and all other tokens with O (Outside or Other). In this case, the class of the NE is also mentioned in label I as shown in Table 1.

Table 1: IO notation for the sentence *Meu pai é Gabriel Raimundo da Silva* (My father is Gabriel Raimundo da Silva)

(Token IO-Notation)
(Meu O) (pai O) (é O) (Gabriel I-PERSON) (Raimundo I-PERSON) (da I-PERSON) (Silva I-PERSON) (. O)

Next, several features are added for each token of the files. These features are used during supervised learning of the CRF prediction model. The features used were the same proposed by (Amaral, 2013), in addition to that feature corresponding to the label assigned by LG. The feature set is presented in Table 2.

The POS-Tagging of a word corresponds to its grammatical class and it was also assigned by the OpenNLP library. When a word does not have one of the previous words (p-1 or p-2) or posterior (p+1 or p+2), the corresponding feature values are ”null”. Table 3 presents an example of a vector of features.

The methodology used for testing is similar. The difference is that 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.

3.1. Local Grammars (LG)

An LG created in Unitex is represented as a set of one or more graphs. The LG built in this work 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 1.

[1] http://unitexgramlab.org/

[2] http://opennlp.apache.org/

Table 2: Feature set assigned to each token

Features	Description
word	current word (position p)
tag	POS-Tagging of the word corresponding to its grammatical class
cap	if the word is composed of only capital letters, only lowercase or mixed
ini	if the word starts with uppercase, lowercase or symbols
simb	if the word is composed of symbols, digits or letters
prevW, prevT, prevCap	word, tag and cap for the word in position p-1
prev2W, prev2T, prev2Cap	word, tag and cap for the word in position p-2
nextW, nextT, nextCap	word, tag and cap for the word in position p+1
next2W, next2T, next2Cap	word, tag and cap for the word in position p+2
tip	label assigned by LG to the word

Table 3: Example of a vector of features to the Gabriel token in sentence *Meu pai é Gabriel Raimundo da Silva*

Token	Vector of features	IO Notation
Gabriel	word=Gabriel tag=prop cap=maxmin ini=cap simb=alpha prevW=é prevT=v-fin prevCap=min nextW=Raimundo nextT=n nextCap=maxmin prev2W=pai prev2T=n prev2Cap=min next2W=da next2T=v-pcp next2Cap=min tip=I-PERSON	I-PERSON

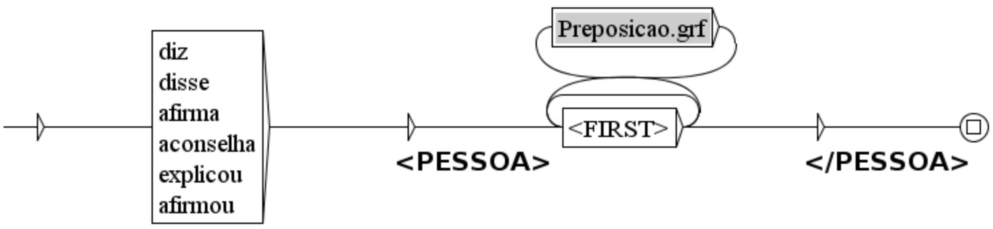

Figure 1: Example of rule in the graph that recognizes the Person category

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 (Lafferty et al., 2001). 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 (Sutton and McCallum, 2012), a linear-chain CRF is a conditional distribution that takes the form shown in Equation 1:

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

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

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

$F = \{f_k(y_t, y_{t-1}, \mathbf{x}_t)\}_{k=1}^{K}$ 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 $\mathbf{x}_t$), its label is Person (y_t) and the previous label (y_{t-1}) is Other and 0 otherwise. The vector $\mathbf{x}_t$ 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[3] 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).

4. Results and Discussion

The annotated files by CRF+LG and NERP-CRF (Amaral, 2013) were submitted to SAHARA[4] for performance evaluation. SAHARA is an online system for automatic evaluations of the HAREM. This system computes metrics of Precision (P), Recall (R) and F-Measure (F) of an NER system after submitting annotated XML files and configures the evaluation desired by the user. In this setting, the evaluation mode is chosen, denoting which task should be evaluated: identification that evaluates only if the recognized string is actually an NE; or classification that in addition to checking the boundaries of the NE also checks if the category is correct. The only modification made to the default configuration was the assignment of the value zero for the β and γ parameters because they correspond to the types and subtypes of the categories that were not classified in this work.

The file annotated by NERP-CRF was obtained as indicated in (Amaral et al., 2014)[5]. We modified the identifiers (ID) of each NE by adding a unique number at the end for the evaluation in the SAHARA due to the NERP-CRF uses the same ID for all NEs in a document and this changes the actual system performance computed by SAHARA. Note that, when a unique ID is not assigned to every NE, the computed metrics do not consider all false positives, only one per document that has false positive. We realized this by studying the evaluation architecture of the Second HAREM and analyzing the files generated by each module.

The results obtained overcome the NERP-CRF results in more than 10% for the Recall metric in the identification task (Table 4) and more than 8% in the classification task (Table 5). For the F-Measure metric, CRF+LG overcome the NERP-CRF results in more than 8% and 7% in the identification and classification tasks respectively, representing considerable gain.

The authors in Santos and Guimaraes (Santos and Guimaraes, 2015) used the GC of the First HAREM as training set and the MiniHAREM as the test set. As (Santos and Guimaraes, 2015) did not present the results for the GC of the Second HAREM, CRF+LG was rerun using the GC that they used for training and testing. The CoNLL-2002[6] script that evaluates the classification task was also used as done by those authors to compute the metrics in our experiments. The selective scenario (categories Person, Place, Organization, Time and Value) was considered in this case because the results presented for the 10 categories of the HAREM were obtained using word-level embeddings previously trained by (Santos and Zadrozny, 2014) who used three other corpus (Portuguese Wikipedia, CETENFolha and CETEMPublico) to perform this unsupervised pre-training. Therefore the comparison with this result would be unfair since the CRF+LG uses only the GC of the First HAREM for the CRF training phase and LG construction. Hence, just for the sake of comparison, the GC of the First HAREM has approximately 78667 words while only the CETEMPublico, one of the three corpus used by (Santos and Guimaraes, 2015), has about 180 million words.

The results are presented in Table 6. Note that CRF+LG achieved a gain of approximately 2% in each metric evaluated.

Table 4: Comparison with NERP-CRF - Identification

Systems	P (%)	R (%)	F(%)
NERP-CRF	73.68	53.79	62.19
CRF+LG	**78.58**	**64.12**	**70.62**

Table 5: Comparison with NERP-CRF - Classification

Systems	P (%)	R (%)	F(%)
NERP-CRF	61.04	43.18	50.57
CRF+LG	**65.46**	**51.75**	**57.8**

Table 6: Comparison with CharWNN

Systems	P (%)	R (%)	F(%)
CharWNN	65.21	52.27	58.03
CRF+LG	**67.09**	**54.85**	**60.36**

We observed some errors when analyzing false positives and false negatives obtained by CRF+LG: prepositions like *de* (of) and conjunctions like *e* (and) that are not considered part of NEs since they are also common outside NEs (e.g., *Joaninha Sampaio* labeled as Person when the name was *Joaninha Sampaio e Melo*); names labeled as Person names when they are part of a larger NE (e.g., *José Mourinho* in *Liderança - As Lições de José Mourinho* that should be labeled as Work); capitalized words labeled as Organization (e.g., *FESTA* which is not NE).

[3] http://mallet.cs.umass.edu/

[4] http://www.linguateca.pt/harem/

[5] http://www.inf.pucrs.br/linatural/recursos_para_reconhecimento_de_entidades_nomeadas/NERP_CRF.xml

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

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.

5. Conclusions

This paper presented a hybrid approach for the Named Entity Recognition in Portuguese texts using Conditional Random Fields and Local Grammars. The term classification obtained initially from LG was sent as a feature for the learning process of the CRF prediction model together with other features. The CRF model performs the final labeling of the NEs. Our approach is a good way to consider the human expertise for capturing the rules that do not appear in examples of the annotated corpus used for training by the CRF.

The results obtained outperform the results of competitive systems reported in the literature when performing under equivalent conditions. It is important to mention that these are the results for a small corpus and the gains can become more expressive when using a larger corpus for training.

In future work, we will investigate the impact of some preprocessing decisions on the performance of the CRF and the impact of using the result of other classifiers to inform new features for the CRF learning process rather than an LG.

6. Bibliographical References

Amaral, D. O., Fonseca, E. B., Lopes, L., and Vieira, R. (2014). Comparative analysis of portuguese named entities recognition tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2554–2558.

Amaral, D. O. F. d. (2013). 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, Brazil.

Constant, M. and Tellier, I. (2012). Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. In *8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 646–650.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING*, volume 96, pages 466–471.

Gross, M. (1999). A Bootstrap Method for Constructing Local Grammars. In Neda Bokan, editor, *Proceedings of the Symposium on Contemporary Mathematics*, pages 229–250. University of Belgrad.

Jiang, J. (2012). Information extraction from text. In *Mining text data*, pages 11–41. Springer.

Konkol, M. and Konopík, M. (2015). Segment representations in named entity recognition. In *International Conference on Text, Speech, and Dialogue*, pages 61–70. Springer.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001*, volume 1, pages 282–289.

Linguateca. (2017). `http://www.linguateca.pt/HAREM/`. Accessed 27/01/2017.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367. Association for Computational Linguistics.

Mota, C. and Santos, D. (2008). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM.

Santos, D. and Cardoso, N. (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.

Santos, C. N. d. and Guimaraes, V. (2015). Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entities Workshop, ACL 2015*, pages 25–33.

Santos, C. N. d. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *ICML*, pages 1818–1826.

Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans et al., editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.