Optimization of Portuguese Named Entity Recognition and Classification by combining Local Grammars and Conditional Random Fields trained with Parsed Corpus

Diego Alves[1], Božo Bekavac[1], Marko Tadić[1], and

Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb 10000, Croatia {dfvalio,bbekavac,marko.tadic}@ffzg.hr,

Abstract. This article presents the results of the study concerning named-entity recognition and classification for Portuguese focusing on temporal expressions. We have used Conditional Random Fields (CRF) probabilistic method and features coming from automatically annotated parsed corpus and local grammars. We were able to notice that Part-of-Speech (PoS) tags are the most relevant information coming from a parsed corpus to be used as feature for this task. No positive synergy emerges from the association of these tags with other linguistic information from the parsed corpus. NooJ local grammar, created to recognize "Time" category entities (without detailing types and subtypes) provides information that surpasses PoS tags as a feature for CRF training in terms of precision and recall. The combination of PoS and NooJ annotations does not bring any advantage.

Keywords: named-entity · conditional random field · portuguese

1 Introduction

Named Entity Recognition and Classification (NERC) involves recognizing information units such as person, organization, location (sometimes), time expressions and others present inside unstructured texts. This Information Extraction task is crucial for many natural language processing (NLP) applications such as question and answering systems and entity-oriented search [4]. The awareness of NERC evaluation started with the Message Understanding Conferences (MUC), more precisely MUC-6 [3], and since then, many different evaluation campaigns have been developed worldwide.

For Portuguese language (both native and Brazilian), the HAREM evaluation campaigns [1] (Avaliação de Reconhecedores de Entidades Mencionadas) [1] established a complex, and therefore, ample set of tags for this task and evaluated several NERC systems based on either machine learning approaches or hand-coded rules in combination with dictionaries, gazetteers, and ontologies. HAREM hierarchy is composed by three levels: the first one composed by ten categories

[1] Translation: "Evaluation of Recognizers of Mentioned Entities".

(”Person”, ”Place”, ”Organization”, ”Time”, ”Work”, ”Thing”, ”Event”, ”Miscellaneous”, ”Abstraction” and ”Value”), the second level concerning 36 types and the third one corresponding to 21 sub-types.

A rule-based model for Portuguese NERC has been proposed inside Port4NooJ v3.0 [8], however, the evaluation of this tool has not been provided and the module is not fully available. Pirovani, J. and De Oliveira, E. proposed in their article [9] the association of local grammars with Conditional Random Fields (CRF) probabilistic method based on a training set containing Part-of-Speech (PoS) tags achieving better results than previous works concerning HAREM. In this study, the authors considered only the following HAREM named-entity categories: ”Person”, ”Place”, ”Organization”, ”Time” and ”Value”. Types and sub-types of these categories (corresponding to second and third level of the HAREM hierarchy) were not taken into consideration by them. Precise evaluation of the deeper structure of HAREM is not provided in previous studies, like in the proposed method using Stencil and NooJ [7] which considers only ”Date”, ”Hour” and ”Period” types.

Conditional Random Fields [6] is a machine learning method, which have been widely used in several types of NLP tasks, including NERC. This task is treated as a sequence labeling problem and a conditional model is built from a training set to predict which is the most appropriate labeling sequence given an input sentence.

Local Grammars correspond to the representation of contextual rules. They are finite-state grammars or finite-state automates that represent sets of utterances of a natural language [2]. NooJ software allows the development of formalized descriptions of natural languages in terms of electronic dictionaries and grammars represented by organized sets of graphs [12].

The IOB format, proposed by Ramshaw and Marcus [11], which allows a more detailed analysis of the boundary identification of the entities, was chosen for our study following the work of Pirovani and De Oliveira [9].

Our proposal is to focus on the “Tempo” (”Time”) annotation. Our hypothesis is that by combining linguistic information coming from a parsed corpus with pre-annotations in terms of Named-Entity Recognition provided by NooJ local grammars, we can improve precision and recall considering the whole complexity of the ”Time” category described in HAREM. By understanding and identifying possible synergies between different features to be used in CRF models training, these findings can be used in any other NERC CRF system to enhance final evaluation metrics of this task.

This paper is organized in 6 sections. Section 2 details the HAREM dataset used in our study, then Section 3 presents the 3-step methodology adopted in this work. The obtained results are presented in Section 4 and discussed in Section 5. Section 6 presents conclusions and perspectives for future work.

2 Data Description

The golden collection used in our study (xml file) was created for the second edition of the HAREM evaluation campaign for Portuguese [2], addressing named entity recognition [1]. It is composed by 129 documents from different domains: news, didactic, opinion, blog, questions, interview, legal, literary, promotional and private manuscripts. The corpus was manually annotated according to specific guidelines and contains 4053 sentences and 89634 tokens.

HAREM named-entity hierarchy is composed by 3 levels:

- First level: 10 categories.
- Second level: 36 types.
- Third level: 21 sub-types.

As previously mentioned, our focus is on the "Time" category which is described bellow:

Table 1. Description of "Time" category from HAREM named-entity hierarchy in terms of types, sub-types and number of occurences in the golden collection.

Type	Sub-type	Number of occurrences
TIME_CALENDAR	DATE	873
	HOUR	37
	INTERVAL	63
DURATION	-	56
FREQUENCY	-	71
DURATION	-	89

In total, there are 1189 occurrences of "Time" entities in the selected corpus, 73% correspond to the type "TIME_CALENDAR" and sub-type "HOUR".

This golden collection was preprocessed by two python scripts:

- Format change: from inline XML Schema to IOB format.
- Random selection of sentences to compose train and test sets respecting the ratio 70/30.

Therefore, we have established a training set composed by 2842 sentences, 63032 tokens and 808 named-entities, and a test set containing 1211 sentences, 26602 tokens and 352 entities.

3 3-step Experimental Design

We have developed a 3-step methodology to analyse the influence of different features and their synergy when training CRF models for the task Named-Entity

[2] Corpus file called "CDSegundoHAREMReRelEM" and available at: https:www.linguateca.pt

Recognition and Classification. The first step consisted in training CRF models using Part-of-Speech, Morphological and Dependency Parsing tags as features. In the second step, NooJ software was used to generate a local grammar to identify and annotate named-entities of the "Time" category. In the last step, we have merged features coming from steps 1 and 2 to analyse their combined influence when training CRF models. Each step will be detailed in the following sub-sections.

3.1 Step 1: CRF trained with Parsed corpus

As mentioned above, in this step, linguistic information coming from a parsed corpus is used as features to train CRF models. The objective was to identify the best combination that allows to enhance the metrics results (precision, recall and F1-measure) for the NERC task.

The golden collection provided by HAREM does not contain any linguistic information beside the identification of named-entities. Therefore, to provide additional features, we have parsed the train and test sets using UDpipe tool v.1.2.0 [14] with an available Portuguese model (BOSQUE v.2.4), which was developed using Bosque treebank [10].

We have used sklearn_crsuite v.0.3.6 python library (with Python 3.6.2) [5] which allows the implementation of Conditional Random Field models [15] for labeling sequential data.

The basic features that were used in all our CRF tests and which consists our baseline (simplest model, with lesser features) are:

- Lower case: if token is composed only by lower case characters.
- Upper case: if token is composed only by capital letters.
- Title: if token starts with a capital letter followed only by lower characters.
- Digit: if token is composed by digits,

Each token (n) also receives information concerning the previous (n-1) and following (n+1) tokens.

Hence, the following combination of features have been tested:

Table 2. Different combination of features coming from parsed corpus used to train CRF models.

Test number	Features
1	Basic Features (baseline)
2	Basic Features + Part-of-Speech tags
3	Basic Features + Dependency Parsing labels
4	Basic Features + Morphosyntactic tags
5	Basic Features + Part-of-Speech + Dependency Parsing
6	Basic Features + Part-of-Speech + Morphosyntactic
7	Basic Features + Part-of-Speech + Morphosytactic + Dependency Parsing

3.2 Step 2: Development of Local grammar to identify ”Time” entities

The main idea of this step is to develop a local grammar with NooJ software v.6.1 [12] and the general dictionary provided by the Port4NooJ resource [8] to identify ”Time” expressions in the train and test sets. Our aim is to give this additional information as a feature when training CRF models.

The local grammar that has been created is composed of 22 graphs and allowed us to recognize and annotate entities concerning the category level. We have not considered all the types and sub-types of ”Time” in this step.

NooJ xml tags have been transformed into IOB format using a python script.

3.3 Step 3: Combination of features coming from parsed corpus and local grammar

In this final step, we have combined the features identified in the first step providing the best results in terms of named-entity recognition and classification with the pre-annotations performed by the local grammar generated using NooJ software.

Table 3. Different combination of features coming from parsed corpus and NooJ local grammar to train CRF models.

Test number	Features
8	Basic Features + NooJ local grammar ”Time” tags
9	Basic Features + Part-of-Speech + NooJ local grammar ”Time” tags

4 Results

4.1 Step 1: CRF trained with Parsed corpus

In the following table, we present the results of the tests that were conducted using CRF models trained with linguistic information coming from HAREM golden set corpus parsed with UDpipe tool.

The best results in terms of precision, recall and F1-measure are obtained using the basic features associated with Part-of-Speech information, with an increase in these metrics compared to the baseline. Dependency Parsing labels allows an improvement in terms of recall but not of precision compared to the baseline. Morphosyntactic tags used as a feature increase precision but has no effect in terms of recall. The association of Part-of-Speech tags with the other linguistic data coming from the parsed corpus (tests 5, 6 and 7) does not have any positive impact in the metrics compared to the results obtained using Part-of-Speech alone.

Table 4. Results of different combination of features coming from parsed corpus used to train CRF models in terms of Precision (P), Recall (R) and F1-measure (F1).

Test number	Features	P	R	F1
1	Basic Features (baseline)	0.809	0.636	0.700
2	Basic Features + PoS	0.838	0.671	0.735
3	Basic Features + Dep. Parsing	0.805	0.664	0.723
4	Basic Features + Morphosyntactic	0.823	0.634	0.708
5	Basic Features + PoS + Dep. Parsing	0.807	0.670	0.727
6	Basic Features + PoS + Morpho	0.830	0.658	0.727
7	Basic Features + PoS + Morpho + Dep. Parsing	0.823	0.655	0.721

4.2 Step 2: Development of Local grammar to identify ”Time” entities

A local grammar composed by 22 graphs has been created to identify entities concerning ”Time” category. Each token from train and test sets were annotated with one of the following tags: ”B-Time”, ”I-Time” or ”O”.

The accuracy of the annotations provided in this step is presented in the following table. We have examined both train and test sets.

Table 5. Evaluation of annotations from NooJ local grammar in terms of Precision (P), Recall (R) and F1-measure (F1).

Data-set	P	R	F1
Train	0.870	0.661	0.751
Test	0.847	0.660	0.741

Precision and recall values for train and test sets are relatively similar. These results show that the local grammar has been built favoring precision over recall. This local grammar can be improved by increasing the complexity of “Time” expressions that it can identify and by the creation of disambiguation grammars that would allow better identification of entities containing ambiguous words such as “era” which in Portuguese can correspond to “era” as in English and also the past tense of the verb “to be”.

4.3 Step 3: Combination of features coming from parsed corpus and local grammar

In this final step, we have analysed the combination of the best model coming from training of CRF models using parsed data with the annotations provided by NooJ as an additional feature. Results are presented on the table below:

It is possible to notice that the combination of NooJ annotations as a feature with the basic ones enables the CRF model to achieve better results than the best

Table 6. Results of the combination of features coming from parsed corpus and NooJ annotations to train CRF models in terms of Precision (P), Recall (R) and F1-measure (F1).

Test number	Features	P	R	F1
1	Basic Features (baseline)	0.809	0.636	0.700
2	Basic Features + PoS	0.838	0.671	0.735
8	Basic Features + NooJ annotations	0.867	0.725	0.757
9	Basic Features + PoS + NooJ annotations	0.797	0.675	0.700

model obtained in the first step (using Part-of-Speech tags). NooJ “Time” expressions identification allows us to increase precision by approximately 6 points while recall is increased by almost 9 points. When NooJ information is combined with Part-of-Speech tags, precision is close to the value obtained using only the basic features (baseline) and there is only a small increase in recall. Therefore, combining basic features with NooJ annotations provides the best CRF model for NERC of temporal expressions in Portuguese, using HAREM golden set.

As presented in table 1, HAREM proposes a classification of ”Time” expressions composed by types and sub-types. For further understanding on how the best model works, it is important to analyse the results for each possible tag in detail.

Table 7. Results for each possible named-entity tag using CRF model trained with basic features and NooJ annotations in terms of Precision (P), Recall (R) and F1-measure (F1).

Test number	P	R	F1
B-DURATION	0.333	0.059	0.100
I-DURATION	0.333	0.089	0.140
B-FREQUENCY	1.000	0.545	0.706
I-FREQUENCY	1.000	0.516	0.681
B-GENERIC	0.500	0.138	0.700
I-GENERIC	0.800	0.082	0.148
B-TIME_CALENDAR-DATE	0.841	0.821	0.831
I-TIME_CALENDAR-DATE	0.828	0.857	0.842
B-TIME_CALENDAR-HOUR	1.000	0.167	0.286
I-TIME_CALENDAR-HOUR	0.750	0.194	0.308
B-TIME_CALENDAR-INTERVAL	0.778	0.438	0.560
I-TIME_CALENDAR-INTERVAL	0.778	0.404	0.532

”Duration” type presents the worst values for both precision and recall. The tag ”B-GENERIC” also has considerable lower values for these two metrics than the other tags. In general, recall results are much lower than precision, except for type ”Time Calendar”, sub-type ”Date”, for which the values are almost similar. These instances are the most numerous in the golden set, corresponding to 73% of all named-entity occurrences. Only the types "Time Calendar", subtype "Date" and "Frequency" have recall results higher than 0.500.

5 Discussion

By analysing the results presented in the previous section, it is noticeable that when using information from an annotated parsed corpus, the most relevant linguistic information to be used as a feature for CRF training is Part-of-Speech. When used alone, Part-of-Speech increases the values of all metrics. The association with other information coming from the parsed corpus showed no further increase, thus, no positive synergy has been identified. This observation corroborates with what is usually seen in the literature concerning CRF for NERC task where, usually, only PoS is used as feature [9].

We have used UDpipe tool (with BOSQUE model) to annotate HAREM golden set. As this task was automatically done, some bias is introduced in our study. According to UDpipe official website, when evaluated with their test set, BOSQUE model achieved Part-of-Speech and Morphosyntactic tagging accuracy higher than 90% while dependency parsing tagging (LAS metric) of 85.65% [13].

Table 6 shows that when using CRF method for this specific task, generating a local grammar has better final results than using Part-of-Speech as a feature. Also, no positive impact has been observed when associating both PoS and NooJ annotations. We have chosen to focus on temporal expressions for this study, thus, it seems pertinent to proceed with the same analysis for other main categories present in HAREM to validate these observations.

Therefore, the hypothesis that the best results for NERC can be achieved by the combination of linguistic features from parsed corpus combined with annotations from local grammars is not confirmed. No positive synergy has been observed between them. However, it was possible notice that using only the local grammars annotations as feature for CRF training increases considerably the overall results for this task.

The evaluation of the local grammar showed that it can still be improved, specially in terms of recall, which would probably lead to an enhancement of the CRF model . Other possible improvement would be detailing, in this grammar, types and sub-types of HAREM, focusing on the cases with low recall values. It is possible to see in table 7 that entities corresponding to type "Time Calendar", sub-type "Date" are the ones with highest value of recall, which probably influenced by the number of occurrences in the golden set (much higher compared to other entities). Thus, it is important for further development of the local grammar to focus on what is least represented in HAREM.

Pirovani, J. and De Oliveira, E. [9] also used local grammar and PoS annotations associated with CRF models for NERC task, however, no proof of the synergy between these two types of features was demonstrated. Also, the highest F1 measure obtained by the proposed system (for all HAREM categories, not considering second and third levels of the hierarchy) was 60,4%, while our system achieves 75,7% for the "Time" category considering all its types and sub-types. Additionally, the best NERC system identified during the Second HAREM evaluation campaign [1] presents a F1 measure slightly inferior to 60%. Our system composed only by annotations from local grammar and basic CRF features seems, therefore, a great improvement for this task. Nevertheless, as mention above, since our focus was on the recognition and identification of the detailed ”Time” category structure only, further analysis considering the whole HAREM structure must be conducted.

6 Conclusions and Future Directions

We have presented a detailed study with the aim to analyse the influence of different features when training CRF models to perform the task of named-entity recognition and classification for Portuguese. Our focus was on temporal expressions, using the hierarchy established by HAREM initiative [1].

First, we have identified that among all possible features that can be used from a parsed corpus, the most relevant one for this specific task is Part-of-Speech tags used alone. Associating different features did not provide better results compared to the model trained only with PoS.

Second, we have built a local grammar capable of identifying ”Time” category entities using NooJ software to use this information as a feature for CRF training. The evaluation of this grammar showed that it has room to improvement, specially in terms of recall.

In the final step, we observed that NooJ annotations used as a feature enabled us to create the best CRF model in terms of precision and recall. No positive synergy has been observed when NooJ annotations were combined with PoS tags. Thus, it seems more pertinent to generate local grammars that recognizes entities and their boundaries than to use an automatic tool to parse the corpus. Compared to previous works, it is possible to observe that the overall metrics of our system, even though focusing only on ”Time” category, are better.

Therefore, we can conclude that the combination of local grammar information with linguistic annotations as CRF features does not necessarily improve overall results. Previous studies did not evaluate this synergy. Local grammars alone provide sufficient information to increase precision and recall for NERC task.

As next steps, it would be interesting to proceed with the improvement of the local grammar, increasing the recall and introducing the whole complexity of ”Time” expressions as described in HAREM guidelines, considering all types and sub-types. Also, it would be relevant to test other machine learning methods to verify if the lack of synergy observed here is due to the algorithm tested. And, finally, it seems pertinent to complete this study by considering the whole HAREM hierarchy to be able to compare precisely the overall results of our system with the ones proposed in the literature.

7 Acknowledgements

The work presented in this paper has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no.812997 and under the name CLEOPATRA (Cross-lingual Event-centric Open Analytics Research Academy).

References

1. Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., Santos, D.: Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In: quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association. European Language Resources Association (2010)
2. Gross, M.: A Bootstrap Method for Constructing Local Grammars. In: Bokan, N. (ed.) Proceedings of the Symposium on Contemporary Mathematics, pp. 229–250. University of Belgrad (1999), https://halshs.archives-ouvertes.fr/halshs-00278319
3. Hirschman, L.: The evolution of evaluation: lessons from the message understanding conference. Computer Speech and Language pp. 208–305 (1998)
4. Jiang, J.: Information extraction from text. Mining text data pp. 11–47 (2012)
5. Korobov, M.: sklearn-crfsuite 0.3. https://sklearn-crfsuite.readthedocs.io/en/latest/index.html (2015), (Accessed on 20/05/2020)
6. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning. pp. 282–289. Morgan Kaufmann, San Francisco, CA (2001), citeseer.ist.psu.edu/lafferty01conditional.html
7. Mota, C., Silberztein, M.: Em busca da máxima precisão sem almanaques: O stencil/nooj no harem. In: Diana Santos, N.C. (ed.) Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, pp. 191–208 (2007)
8. Mota, C., Carvalho, P., Barreiro, A.: Port4NooJ v3.0: Integrated linguistic resources for Portuguese NLP. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1264–1269. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1201
9. 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). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1705
10. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva, V.: Universal dependencies for portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling). pp. 197–206. Pisa, Italy (September 2017), http://aclweb.org/anthology/W17-6523
11. Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Third Workshop on Very Large Corpora (1995), https://www.aclweb.org/anthology/W95-0107

12. Silberztein, M.: NooJ Manual (01 2003)
13. Straka, M.: UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 197–207. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/K18-2020, https://www.aclweb.org/anthology/K18-2020
14. Straka, M., Straková, J.: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (August 2017), http://www.aclweb.org/anthology/K/K17/K17-3009.pdf
15. Wijffels, J., Okazaki, N.: crfsuite: Conditional random fields for labelling sequential data in natural language processing based on crfsuite:a fast implementation of conditional random fields (crfs) (2007-2018), https://github.com/bnosac/crfsuite, r package version 0.1