Extraction of entities in Portuguese from the Second HAREM Golden Collection

Nicolas Eymael da Silva, Dante Augusto Couto Barone and Eduardo Gabriel Cortes
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
{nesilva, barone, egcortes}@inf.ufrgs.br

***Abstract*—Information Extraction is an essential process for automatically building a Knowledge Graph, a knowledge base representing knowledge through semantic connections, and has been gaining focus in recent years. One of the tasks required during this construction is Named Entity Recognition, responsible for identifying and classifying the entities in the text. This task is the first step to generate the tuples that form the Knowledge Graph. Although there are already works that deal with this task, many of them are focused on the English language and few on Portuguese. Therefore, the goal of this work was the development of machine learning models capable of extracting entities from texts in Portuguese. The model was used to extract entities through Bidirectional Encoder Representations from Transformers and was trained and evaluated using a simplified version of the Second HAREM Golden Collection dataset, a golden standard for NLP in Portuguese. After evaluating the model, it was observed that the results obtained in the Named Entity Recognition task were promising for the primary classes present in the dataset.**

***Index Terms*—Named Entity Recognition, HAREM, Knowledge Graph**

I. Introduction

Information Extraction (IE) is an important task in natural language processing and text mining, which consists of extracting structured information from unstructured or semi-structured texts [1]. This information can be presented to the user or improve other systems, such as search engines. The first IE systems worked using rule-based models. That is, the information was identified using linguistic patterns developed by humans. These systems can achieve good performance, but creating rules manually is laborious and the rules are highly domain-dependent. Because of these limitations, the researchers decided to approach this task through statistical machine learning models.

A task derived from IE is Open Information Extraction (OIE). OIE systems seek to extract all important information, regardless of the domain, from a large and diverse corpus. This information can be helpful to entities and relations, which are usually represented by tuples. Finally, these tuples can be used in the construction of databases, such as a Knowledge Graph (KG) [2]. This KG can then be used in several applications, such as Financial Analytics, Question Answering, and others.

To generate these tuples, Named Entity Recognition (NER) is required. The NER is an IE subtask responsible for identifying and classifying the entities present in the texts. Although many works deal with this task, they usually focus on the English language and not Portuguese [3].

Among the works focused on Portuguese, the one that stands out the most is HAREM. HAREM is an evaluation contest organized by Linguateca, which aims to carry out the evaluation of NER systems for the Portuguese language [4]. The HAREM corpus is a reference in the NLP area of the Portuguese community and is characterized by having a large set of texts annotated and validated by humans.

Therefore, the goal of this work is the development of a computational model capable of extracting entities in the Portuguese language. This model will be trained and evaluated using a simplified version of the Second HAREM corpus. After the executions, the results of the task will be analyzed and compared with other related systems.

The rest of this work is organized as follows. Section II provides an overview of related work. Section III presents the procedures and tools adopted at each stage of this work. Section IV presents an analysis of the results obtained in the experiments. Finally, in section V the conclusions are presented.

II. Related Work

In this section, the works related to the NER task using the Second HAREM are presented. The research of the works was carried out through Google Scholar. The methods and results of each study are described below.

At the Second HAREM Workshop, organized by Linguateca in 2008, 10 systems were participating in the identification and classification of named entities [4]. Each participant performed 1 to 4 runs with different scenarios. The purpose of identifying entities is only to find the entities present in the text, while the classification of entities also involves finding out in which category the entity fits.

Among all the participants, the Priberam [5] and REMBRANDT [6] systems are the ones that obtained the best performances, taking into account both tasks. Both the Priberam system and the REMBRANDT and most of the participating systems use a set of rules and clauses generated manually in combination with dictionaries and ontologies.

In addition to the systems participating in the Workshop, other NER models also used the Second HAREM dataset in the following years.

The NERP-CRF system [7], unlike the systems mentioned above, uses a machine learning model based on Conditional Random Fields (CRF) for the task of entity classification. The NERP-CRF showed better results than the participating systems for precision metric (83.48%) and F-score (57.92%).

However, this result is biased because a single corpus, the golden collection (GC) of the Second HAREM, was used for training and test through cross-validation. A second run was performed using the First HAREM GC as training and the Second HAREM GC as a test to make the comparison fairer with the participating systems. This second run showed slightly worse results than the first, with an accuracy of 80.77% and an F-score of 48.43%.

The CRF+LG system [8] is a hybrid system that uses linguistic methods and machine learning approaches in the NER task. The CRF+LG combines labeling obtained by Conditional Random Fields (CRF) with a term classification obtained from Local Grammars (LGs). The experiments were performed using the First HAREM GC for training and the Second HAREM GC for testing. The results obtained from the experiments indicate an F-score of 70.62% and 57.8% in the identification and classification of entities, respectively.

Another system worth mentioning is the BERT-CRF, proposed by [9]. This NER system combines the transfer capabilities of BERT with the structured predictions of CRF. However, the experiments were carried out using only the First HAREM, making it difficult to compare the results with the other works already mentioned.

Table I summarizes the relevant information on the systems related to the entity classification task, while Table II presents the metrics obtained for the entity identification task. The NERP-CRF system did not provide metrics for the task of identifying entities.

TABLE I
NER SYSTEMS AND RESULTS OF ENTITY CLASSIFICATION.

System	*Year*	*Precision*	*Recall*	*F-score*
Priberam	2008	64.17%	51.46%	57.11%
REMBRANDT	2008	64.97%	50.36%	56.74%
NERP-CRF (v1)	2014	**83.48%**	44.35%	**57.92%**
NERP-CRF (v2)	2014	80.77%	34.59%	48.43%
CRF+LG	2018	65.46%	**51.75%**	57.8%

TABLE II
NER SYSTEMS AND RESULTS OF ENTITY IDENTIFICATION.

System	*Year*	*Precision*	*Recall*	*F-score*
Priberam	2008	69.94%	**72.29%**	**71.10%**
REMBRANDT	2008	75.77%	62.14%	68.28%
CRF+LG	2018	**78.58%**	64.12%	70.62%

III. METHODOLOGY

This section presents the procedures adopted. Subsection III-A presents the dataset used in this work and the preprocessing steps performed on it, while subsection III-B describes the experiments related to the NER task.

A. Dataset and Preprocessing

The dataset used for both training and testing the system was the Second HAREM GC with manually annotated relations. It is important to note that this dataset was made available by Linguateca in April 2010, so it is not the same as the one used at the Second HAREM Workshop in September 2008.

Two years after the Workshop, Linguateca made a new version of the Second HAREM GC that includes annotating all existing relations between entities throughout the dataset. The distribution of the 7846 registered entities can be seen in Table III. It is important to note that some entities are classified in more than one category.

TABLE III
DISTRIBUTION OF ENTITIES IN THE DATASET. SOME ENTITIES HAVE MORE THAN ONE CATEGORY.

Entity category	*#*
ABSTRACCAO	439
ACONTECIMENTO	368
COISA	388
LOCAL	1608
OBRA	552
ORGANIZACAO	1260
OUTRO	112
PESSOA	2240
TEMPO	1206
VALOR	356

The file made available by Linguateca is in XML format in which each entity has an EM tag and can have a category (CATEG), type (TYPE), and subtype (SUBTYPE), all of which are optional. Due to the complexity of classifying each entity by category, type, and subtype, a new classification system was created with a reduced set of classes. The Beautiful Soup and pandas libraries were used to convert the XML file to a Dataframe to facilitate data manipulation.

Table IV shows the distribution of the new inferred classes. Most of the rules used just directly mapped the old category to the new class, except for INDIVIDUO, ORGANIZACAO, and OUTRO. The total of 7817 entities instead of 7846 is because the entities that did not have the CATEG attribute were discarded.

TABLE IV
NEW DISTRIBUTION OF ENTITIES IN THE DATASET.

Entity type	*#*
ABSTRACCAO	292
ACONTECIMENTO	323
INDIVIDUO	1774
LOCAL	1539
OBRA	489
ORGANIZACAO	1459
OUTRO	390
TEMPO	1199
VALOR	352
TOTAL	7817

B. NER task

The NER task was performed using the Simple Transformers library. For that, it was necessary to modify the dataset to an input format compatible with the library. In this task, two executions were carried out: one to identify which words are entities and the other to classify each one of these entities.

1) Data format: The input data to the ST NER task can be a path to a text file containing the data. When using text files as input, the data should be in the CoNLL format and tagged with the BIO2 format. The CoNLL format is a text file with one word per line with sentences separated by an empty line. The first column in a line should be the word and the second column should be the label.

BIO2 is a tagging format that uses prefixes to classify labels. When an entity appears, the word is marked with a label that begins with the prefix "B-" followed by the entity class. If the entity comprises more than one word, the following labels have the prefix "I-". The words that are not entities are marked with the label "O".

2) Experiments: We use the Simple Transformers model, a Python library developed by Thilina Rajapakse, to facilitate the use of Transformer models, which are state-of-the-art NLP systems that use deep learning models and adopt the mechanism of attention [10]. The ST is based on the Transformers library provided by the Hugging Face community.

After choosing the model type, it is possible to specify the exact architecture and trained weights through pretrained models. These models may be available directly through Hugging Face or through other community contributors.

With the model properly initialized, it is possible to train and evaluate the model using input data. It can be done by separating the input data into training and test sets with cross-validation. After evaluating the model, the metrics and predictions obtained can be analyzed.

The experiments were performed using Google Colaboratory (also known as Colab), a Jupyter notebook environment that runs in the cloud. To decrease the task execution time, the environment was configured to use a GPU.

Two sets of inputs in CoNLL format were generated from the dataset. The first version contains only labels "B", "I", and "O", and was used to identify the occurrences of the entities. The second version was used to classify the entities, so the labels that marked the entities contained the class along with the prefix.

The k-fold cross-validation technique was applied to both input versions. The number made the separation of the folds of documents in the dataset, that is, as the chosen k value was 10 and the dataset has 129 documents, each fold was composed of 13 documents, except for the last fold that was left with 12. For each of the 10 iterations, a NER model was trained using 9 folds and evaluated using the remaining fold. The result of the evaluation was saved for later analysis.

The model used in both executions was the BERTimbau Base [11], a pretrained BERT model for Brazilian Portuguese developed by NeuralMind. For fine-tuning, it was used only one epoch with a learning rate equal to $4e{-}5$.

In addition to configuring the labels present in each version, the "max_seq_length" parameter has also been changed. Since the maximum value for this parameter was 512, the model truncated sentences if they exceeded 512 words. The solution found to work around this problem was to split the sentence in two when it became too long. For the rest of the parameters, the defaults of the NERArgs class were maintained.

IV. Results and Analysis

As already mentioned, the NER task used the Simple Transformers library in all executions. One of the available outputs from the library's NERModel was a list of all the labels that were predicted from the input dataset. These predictions were compared directly with the input labels using the scikit-learn library, a Python ML library that already has methods for extracting metrics. The execution time with all folds of the cross-validation was about 25 minutes, both in the task of identification and classification of entities.

The first results analyzed were from the task of identifying entities, that is, the labels used were only "B", "I", and "O", without the class information. Table V presents the precision, recall, and F-score metrics, both at MACRO-level and MICRO-level, obtained when performing entity identification. It is important to note that, in the case of multiclass classification, the MICRO metrics are always the same, since, for each false positive, there will always be a false negative and vice versa. The results obtained in this task were very encouraging.

TABLE V
MACRO and MICRO metrics from the entity identification task.

		MACRO			MICRO
Fold	*# words*	*P*	*R*	*F*	*P=R=F*
0	9027	95.49%	94.93%	95.20%	98.19%
1	10439	95.12%	95.14%	95.13%	97.27%
2	8695	95.55%	96.74%	96.12%	98.14%
3	3415	94.82%	94.14%	94.48%	96.89%
4	3350	97.48%	96.59%	**97.03%**	98.50%
5	5888	94.52%	95.47%	94.99%	98.01%
6	6382	96.65%	**96.98%**	96.80%	98.43%
7	8238	92.96%	95.38%	94.14%	97.83%
8	16241	**97.54%**	96.51%	97.01%	**98.52%**
9	5755	93.35%	96.05%	94.64%	96.49%
	Avg.	95.35%	95.79%	95.55%	97.83%

The next step was to analyze the results of the entity classification task. In this case, the labels of the input files contained the entity class in combination with some BIO prefix. The metrics obtained in this execution can be seen in Table VI.

Two observations can be made from these values. The first one is related to the discrepancy of MACRO metrics between classification and identification tasks. The F-score, for example, had a difference of approximately 34% between the two tasks. This difference shows that classifying entities is much more complex than just identifying them.

TABLE VI
MACRO AND MICRO METRICS FROM THE ENTITY CLASSIFICATION TASK.

		MACRO			MICRO
Fold	*# words*	*P*	*R*	*F*	*P=R=F*
0	9027	60.27%	60.21%	58.41%	95.31%
1	10439	62.07%	60.57%	60.27%	93.18%
2	8695	61.81%	61.35%	60.52%	95.13%
3	3415	**68.85%**	68.11%	**67.52%**	94.41%
4	3350	62.98%	58.42%	57.98%	93.49%
5	5888	67.25%	**69.64%**	66.05%	**96.38%**
6	6382	63.96%	63.95%	62.51%	94.31%
7	8238	65.03%	63.36%	62.87%	95.26%
8	16241	63.58%	55.92%	53.05%	92.75%
9	5755	66.29%	63.76%	62.04%	91.21%
	Avg.	64,21%	62,53%	61,12%	94,14%

The second observation refers to the large difference between the MACRO and MICRO metrics (about 30%). This is because the model presents a better performance for certain classes of entities. Tables VII and VIII show the normalized confusion matrices for fold 0 and fold 5 created during cross-validation, respectively. All values in the same row of the matrix refer to the ground truth and all values in the same column are the predictions.

When analyzing the matrices, it is possible to observe that the classes "individuo" and "organização" showed good results. Meanwhile, the model has rarely been able to predict entities of type "abstraccao" and "outro". One of the reasons for this disparity in the results is related to the number of instances of each class. While the classes "individuo" and "organização" have about 1500 instances, the classes "abstraccao" and "outro" have 292 and 390 instances, respectively.

Another factor is related to the meanings of the entities. While individuals and organizations represent concrete entities, such as "Bill Gates" or "Microsoft", the entities "other" and "abstraction" present more abstract concepts, such as "Portuguese Language" or "Minimum Wage".

Moreover, since the class "other" includes all entities that could not be classified in the other classes, it ends up becoming complex with a variety of entities. These characteristics end up influencing the performance of the system since the model has more difficulty in detecting patterns in these classes. Another important detail is that the predictions for the label "O" showed a high hit rate in all folds, which widens this difference between the MACRO and MICRO metrics.

After comparing the value of the MACRO F-score with the values of the other systems studied, the obtained F-score was slightly better. While systems like Priberam and CRF+LG have an F-score of 57.11% and 57.8%, respectively, the model proposed in this work presents an F-score of 61.12%. However, these values cannot be compared directly, as the datasets are not equivalent. Several changes were made to the dataset during the course of this work, and even if there were no such changes, the original dataset itself was already different. The Segundo HAREM GC used in this work was an updated version of the collection used in the Workshop.

V. CONCLUSION

Named Entity Recognition task is essential in the process of extracting information from texts. This information can be in tuples, with each tuple containing a relation between entities present in the texts. From this, it is possible to build a database, such as a Knowledge Graph, a set of tuples, which can be used in applications such as Question Answering systems.

However, most of the NER models found in the literature are focused on the English language. Moreover, many Portuguese models adopt a rule-based approach instead of taking advantage of machine learning models. Therefore, the purpose of this work was the elaboration of a system capable of extracting entities in Portuguese using machine learning.

The first step of this work was to study the concepts and related works. Afterward, research was done about which programming language would be used during development, mainly considering the available libraries and which dataset would be chosen to feed the system. Finally, an analysis of the results obtained in the experiments was made.

The system was trained and evaluated using a version of the Second HAREM Golden Collection dataset with a reduced number of classes. After analyzing the results, it was observed that the metrics obtained were below the metrics found in the literature. Since the dataset used in this work was not the same as the one used in related works, it is not possible to make a direct comparison with the results obtained in the literature. Therefore, it would be interesting to adapt the works for the same dataset in the future.

Based on the results of this work, it was concluded that the extraction method was not the most appropriate since the results did not reach expectations. Because of this, the automated construction of KG was not carried out since the tuples that the system would generate would not be sufficiently reliable for that KG to be used safely.

Future work would be interesting to develop new models to obtain better performances in the task. This can be achieved through a more in-depth study of the parameters present in the Simple Transformers library. Another possibility would be to use other NLP libraries, such as spaCy and Transformers. Both libraries are more complex than those used in this work, but they appear to be more effective for these tasks.

REFERENCES

[1] J. Jiang, "Information extraction from text," in *Mining text data*. Springer, 2012, pp. 11–41.

[2] I. Muhammad, A. Kearney, C. Gamble, F. Coenen, and P. Williamson, "Open information extraction for knowledge graph construction," in *International Conference on Database and Expert Systems Applications*. Springer, 2020, pp. 103–113.

[3] P. V. Q. de Castro, N. F. F. da Silva, and A. da Silva Soares, "Portuguese named entity recognition using lstm-crf," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2018, pp. 83–92.

[4] C. Mota and D. Santos, "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem," 2008.

[5] C. Amaral, H. Figueira, A. Mendes, P. Mendes, C. Pinto, and T. Veiga, "Adaptação do sistema de reconhecimento de entidades mencionadas da priberam ao harem," *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca*, 2008.

TABLE VII
CONFUSION MATRIX OF FOLD 0 OF THE ENTITY CLASSIFICATION TASK.

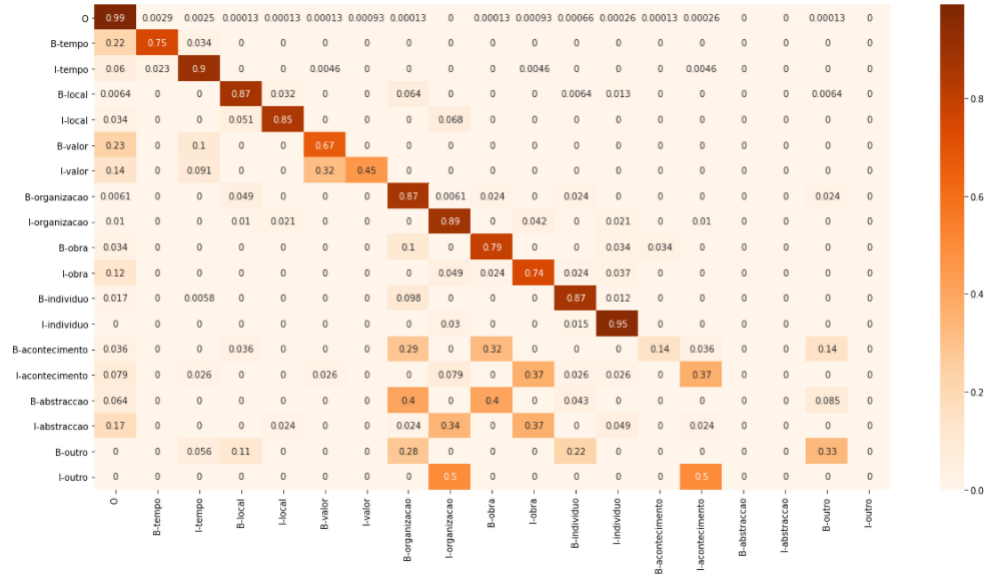
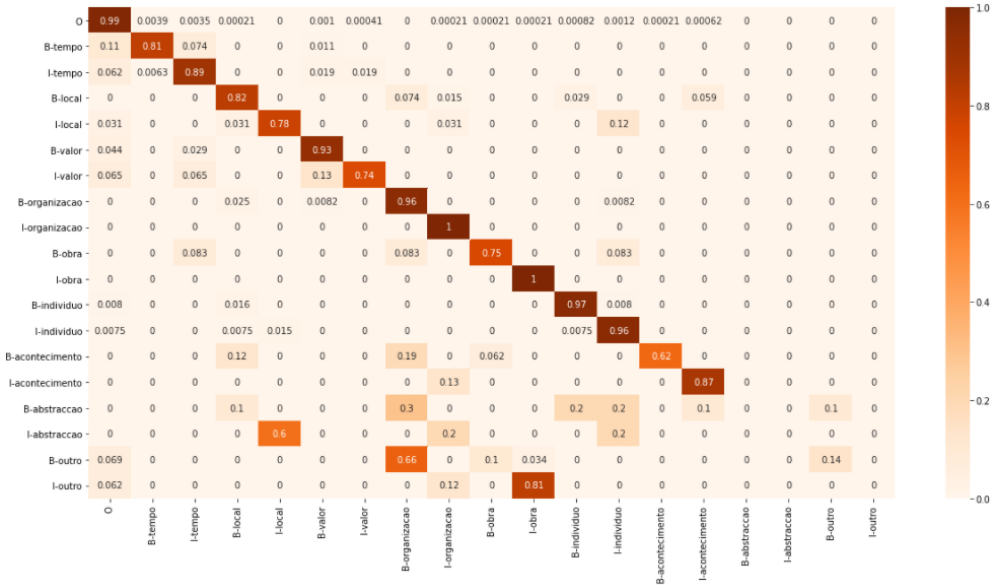


TABLE VIII
CONFUSION MATRIX OF FOLD 5 OF THE ENTITY CLASSIFICATION TASK.



[6] N. Cardoso, "Rembrandt-reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto," *quot; Encontro do Segundo HAREM (Universidade de Aveiro Portugal 7 de Setembro de 2008)*, 2008.

[7] D. O. F. do Amaral and R. Vieira, "Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields." *Linguamática*, vol. 6, no. 1, pp. 41–49, 2014.

[8] J. Pirovani and E. Oliveira, "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.

[9] F. Souza, R. Nogueira, and R. Lotufo, "Portuguese named entity recognition using bert-crf," *arXiv preprint arXiv:1909.10649*, 2019.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[11] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: pretrained BERT models for Brazilian Portuguese," in *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.