Towards Entity Linking, NER in Archival Finding Aids*

Luís Filipe da Costa Cunha[1] and José Carlos Ramalho[2][0000−0002−8574−1574]

[1] University of Minho, Portugal a83099@alunos.uminho.pt
[2] Department of Informatics, University of Minho, Portugal jcr@di.uminho.pt

Abstract. The amount of information present in Portuguese archives has been increasing exponentially over the years. At the moment, the majority of the data is already available to the public in digital format, however, the records are stored as unstructured text, making its data processing challenging.
In this way, it is intended to perform a semantic interpretation of these documents through the identification and classification of Named Entities. For this purpose, the use of Natural Language Processing tools is proposed, training Machine Learning algorithms capable of accurately recognizing entities in this context.
Finally, it is presented a Web platform that implements all the models trained in this paper, as well as some tools that gave support to the entity extraction process.

Keywords: Archival Descriptions · Named Entity Recognition · Machine Learning · Web

1 Introduction

At the moment, in Portugal, there are hundreds of archives spread across the country that keep a diverse universe of archival patrimony in custody. Of these, it is interesting to highlight three archives, the *Arquivo Nacional da Torre do Tombo*, the *Arquivo Distrital da cidade de Coimbra* and the *Arquivo Distrital da cidade de Braga*. These are considered historical archives since they preserve records of various important events that took place throughout the history of the country.

Nowadays, most of the records stored in these archives are already available to the public and can be consulted online via Web portals such as *Digitarq* [1] or *Archeevo* [10]. Despite this, the data provided does not have any kind of annotations being served as natural text, which can cause difficulty in processing and analyzing this type of data.

Thus, it is intended to perform entity recognition in these documents, using Machine Learning(ML) tools, a technique that has been showing excellent results in Natural Language Processing.

In fact, there are already several ML models optimized to extract entities from Portuguese documents, however, the models found were trained in different contexts which means that when applied to archival documents, they reveal results below the intended. Thus, in order to enhance the entity extraction accuracy, new ML models were trained.

Finally, after implementing the entity recognition mechanism, a Web platform was developed and deployed in order to make the generated tools available to the public.

2 Related Work

The study of archive files is something that has been done over the years and of the available computational power is not something new for professional historians. In fact, there are several tools that have been developed over time that assist in the archival data processing.

An example of this is the *HITEX* [13] project, developed by the *Arquivo Distrital de Braga* between 1989 and 1991. This project consisted of semantic model development for the archive historical data, something quite ambitious for that time. Despite this, during its development, it ended up converging to an archival transcription support system, which allowed the transcription of natural text and the annotation by hand of Named Entities enabling the creation of chronological, toponymic and anthroponomic indexes.

Another problem associated with this type of documents was its structure's lack of standardisation. This made it difficult to share information between the archival community both nationally and internationally. To promote data interoperability, in Portugal, guidelines for the archival description have been created that describe rules for standardising the archival descriptions [16]. The purpose of these standards is to create a working tool to be used by the Portuguese archivist community in creating descriptions of the documentation and its entity producer, thus promoting organisation, consistency and ensuring that the created descriptions are in accordance with the associated domain's international standards. The adoption of these guidelines makes it possible to simplify the research or information exchange process, whether at the national or international level.

3 NER Tools

In order to extract entities from archival documents, a subfield of Natural Language Processing was used, Named Entity Recognition. This subject focuses on identifying and classifying Named Entities in text documents, in this case, archival finding aids.

To recognize entities in natural text, one can resort to several mechanisms, such as the simple use of regular expressions, although, some approaches are considered more flexible than others [8]. In fact, nowadays this activity is usually associated with the use of ML tools, which have been showing increasingly accurate results. Initially, Portuguese pre-trained models trained with the *HAREM* [5] and *SIGARRA* [15] datasets were used, however, due to these datasets containing data of a different nature to the context of this paper, the results obtained were below than intended so new training data was generated in order to train new models from scratch.

In this paper, three distinct examples of NER implementations using ML statistical models will be presented, such as the use of different kinds of Neural Networks and the Maximum Entropy algorithm.

3.1 OpenNLP - Maximum Entropy

The first presented tool to perform NER is *Apache OpenNLP* [14] which consists of a machine learning-based toolkit, developed in *Java* programming language, that presents a wide range of ML features for NLP, entity recognition being one of them. To recognize entities in unstructured texts, this tool uses a Maximum Entropy statistical model which, in short, consists of maximizing the entropy of a probabilistic distribution subjected to an N number of constraints [11].

The core of this algorithm is to define a set of features that allow the introduction of known information about the problem domain. Then, the function of Information Entropy is used to maximize the entropy of the models that satisfy the restrictions imposed by the previously selected features, in order to choose the model that makes the smallest implicit assumptions possible.

3.2 spaCy - Convolutional Neural Network

Another tool used to experiment the entity recognition potential in this domain was *spaCy* [17], an open-source library developed in *Python*, for advanced natural language processing.

Again, this tool approaches NER with the use of ML algorithms, this time with the use of Deep Learning, Convolution Neural Networks(CNN).

In fact, the use of Deep Learning in this area is more and more common due to the results that this approach has shown. In this case, *spaCy* uses a transition-based approach [18], i.e., a system that has a set of actions at its disposal, for example, associating an entity label with a certain token or not. Thus, the challenge of this approach is to determine what action to take. For this, a “Deep Learning framework” is implemented, which helps the system to predict the action to be taken, in favor of the Named Entities’ correct identification and classification.

3.3 TensorFlow - BI-LSTM-CRF

The last tool used in this paper was *TensorFlow*, an open-source library focused on ML features that allow to develop and train models in a similar way to the learning method of the human mind. Using this tool, it is intended to create a system capable of recognizing entities in Portuguese archive texts. For this, it was necessary to implement tokenizer mechanisms, create a vocabulary and generate word embeddings from this vocabulary, in order to create and train the NER statistical model.

Usually, *TensorFlow* is associated with the use of Deep Learning, and there is a kind of Neural Networks that is really good at processing sequential data, Recurrent Neural Networks (RNN) [6], which makes them the perfect algorithm for analyzing unstructured text. Despite this, it was previously demonstrated by the research community that this algorithm alone, lacks important features when it comes to NER. Thus, an "upgraded" version of it was used, a Bidirectional Long Short Term Memory (BI-LSTM) with a Cross Random Field (CRF) component on top of it.

In short, an LSTM consists of a Recurrent Neural Network (RNN) to which a memory component has been added, that allows an RNN to be able to preserve Long Term Dependencies along its chain [12]. That said an RNN is unidirectional and, in order to accurately classify a token's label, the model must take into account the context of the token's neighborhood in both directions. Thus, two LSTMS are used, one responsible for the previous and the other for the future context creating a BI-LSTM. Finally, on top of this is added a CRF component that is responsible for encoding the best tagging sequence, boosting the tagging accuracy [9].

Thus, a BI-LSTM-CRF is generated, the model that obtained state-of-the-art results on several NLP tasks in 2015 [7].

4 Models' Results

One of the main objectives of this entity recognition was to optimize the obtained results on a set of metrics: Precision, Recall and F1-Score. Thus, it is necessary to train new models so that the environment in which they are trained is as close as possible to the target context.

In order to train the ML models, it was necessary to create training data associated with the context of the archives documents. Thus, a set of national archives corpus was selected in order to begin the process of annotating a representative fraction of each one. In total, 6302 sentences were annotated which are constituted by more than 160000 tokens. As for entities count, 17279 names of People, 6604 Places, 2980 Dates, 978 Professions or Titles and 843 Organizations were annotated, making a total of 28684 entities.

After being annotated, each dataset was separated into two parts, with 70% of each used for training and 30% for model validation. With the validation and training data ready, the models were trained and then validated. During the training process, individual optimizations for each tool were performed in order to obtain the best possible results, for example, defining hyper-parameters. Then the validation process started and, as can be seen in the Table 1, very satisfactory results were obtained.

In this case, deep learning was clearly a winner in this NLP subfield. In fact, *OpenNLP* achieved lower results than other tools obtaining F1-score values between 69.80% and 99.71% followed by *spaCy* which achieved values between 75.98% and 99.94% and finally *Tensorflow* with the BI-LSTM-CRF model achieving values between 78.89% and 100%.

Corpus	Tool	Precision(%)	Recall(%)	F1-Score(%)
IFIP	OpenNLP	89.43	83.60	86.41
	spaCy	86.99	88.71	87.84
	TensorFlow	92.84	96.85	94.08
Família Araújo Azevedo	OpenNLP	81.94	63.67	71.66
	spaCy	75.19	76.78	75.98
	TensorFlow	78.22	82.47	78.89
Arquivo da Casa Avelar	OpenNLP	88.84	81.68	85.11
	spaCy	87.18	87.18	87.18
	TensorFlow	86.83	92.21	87.99
Inquirições de Genere 1	OpenNLP	99.60	99.53	99.57
	spaCy	98.31	96.74	97.52
	TensorFlow	100	100	100
Inquirições de Genere 2	OpenNLP	74.70	65.61	69.80
	spaCy	79.96	92.21	87.26
	TensorFlow	93.70	98.34	94.82
Paróquia do Jardim do Mar	OpenNLP	99.71	99.71	99.71
	spaCy	99.15	100	99.57
	TensorFlow	100	99.60	99.72
Paróquia do Curral das Freiras	OpenNLP	93.49	99.69	96.49
	spaCy	99.98	99.90	99.94
	TensorFlow	100	100	100

Table 1. Named Entity Recognition Results.

5 Web Platform

Throughout this project, several tools were generated that allowed to facilitate and support its development. In order to encourage the investigation of this area of NLP, applied to archival documents, all the produced material was made public through the creation of a Web platform.

This platform serves as a portfolio of the project, implementing several produced mechanisms with the main objective of allowing its users to take advantage of the ML models generated with the three tools, *OpenNLP*, *spaCy* and *TensorFlow*, enabling the execution of Named Entity Recognition in new unstructured text documents. The purpose of creating this platform is to make available the tools created to the community, which contain the following features:

- Enables users to perform Named Entity Recognition with three different ML statistical models.
- Enables sorting the results by entity type, alphabetical ordering and repeated entities filtering.
- Supports the import of text files as input of the NER ML models.
- Export of extracted entities in CSV and JSON file formats.
- Presents results from previous entities recognition so that it is possible to verify real cases application of each model in several different datasets.
- All annotated datasets are available for download in BIO format.

- Presents various dataset formats that are used in this subfield, such as CSV and BIO providing parses that allow the conversion of datasets between different formats.

It is important to mention that all available ML models on this platform were trained with archival documents, that is, they are expected to be able to perform in similar contexts. In this way, when using these models to recognize entities in documents of a different nature, poor results are expected.

Finally, it is also interesting to mention that, with the use of ML statistical models in archival fonds, it was possible to perform an entity extraction that resulted in hundreds of thousands of extracted named entities. This result can be observed on the platform.

5.1 Implementation

The Web application was designed with micro-services-based architecture and has two micro-services that correspond to the back end, which was developed in Node.js, Python and Java, and the front end, implemented in Vue.js and complemented with the Vuetify framework.

The back end server is responsible for receiving, processing and responding to HTTP requests. In this way a node.js server complemented by the *Express* [3] library , is responsible for managing all API routes and, when necessary, delegates the data processing to the corresponding tools. This happens for example in NER requests, which are processed by the ML models of *OpenNLP*, *spaCy* and *TensorFlow*. In this way, the node.js's *child_process* [2] library is used in order to create child processes that execute programs in Java and Python, waiting for the output of their execution, and then forwarding the response to the client.

On the other hand, the front end was developed with a progressive javascript framework for creating reactive interfaces, Vue.js. This tool is focused on the view layer (client-side). It has a small learning curve so it is fairly approachable, allowing the creation of a performant and maintainable interface due to its reusable components mechanism that allows isolating all logic from the views.

Finally, docker images of the application were created for its deployment, so at the moment, it is hosted on the servers of the Department of Informatics, University of Minho at [4].

6 Conclusion

As demonstrated in the validation of the ML models, this NER technique reveals great potential in this context, obtaining F1-score values greater than 80% in most of the tested corpus. It is also important to note that the algorithms that take advantage of Deep learning obtained better results. Furthermore, analyzing the available results on the Web platform, the trained models were able to extract hundreds of thousands of Named Entities from archival fonds by annotating only a small fraction of them and use that fraction for training the tools.

Thus, it is concluded that the use of ML tools to extract entities from archival documents is a viable approach, and it creates the opportunity to generate different navigation mechanisms and create relations between information records.

7 Future Work

One way to improve the obtained results in entity recognition is to increase the amount of annotated data. In fact, training models with a larger data set makes them able to perform in a wider variety of contexts. Another way to improve the model's results would be to improve the used technologies. The the Attention Mechanism [19] has shown innovative results in this NLP sub-field, so it would be interesting to test this technology in the archive context.

On the other hand, the extracted entities translate into valuable information about their corpus. These data can be explored to implement new tools, for example, Entity Linking mechanisms, enabling navigation between different documents through the relationship between entities.

Finally, in order to complement the created Web platform, it would be interesting to use the trained ML models to create a support tool for unstructured text annotation, taking advantage of Active Learning techniques.

References

1. Arquivo nacional torre do tombo, https://digitarq.arquivos.pt/, accessed in 18-04-2021
2. Node.js v16.4.0 documentation, https://nodejs.org/api/child_process.html, accessed in 17-03-2021
3. Node.js web application framework, https://expressjs.com/, accessed in 10-04-2021
4. Costa Cunha, L.F., Ramalho, J.C.: http://ner.epl.di.uminho.pt/
5. Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P.: Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010), http://www.lrec-conf.org/proceedings/lrec2010/pdf/412_Paper.pdf
6. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (2013). https://doi.org/10.1109/ICASSP.2013.6638947
7. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging (2015)
8. Ingersoll, G.S., Morton, T.S., Farris, A.L.: Taming text: how to find, organize, and manipulate it. Manning, Shelter Island (2013), oCLC: ocn772977853
9. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers (2016). https://doi.org/10.18653/v1/p16-1101
10. Arquivo Regional e Biblioteca Pública da Madeira, A.: https://arquivo-abm.madeira.gov.pt/, accessed in 10-03-2021

11. Maxent, O.: The maximum entropy framework (2008), http://maxent.sourceforge.net/about.html, accessed in 24-09-2020
12. Olah, C.: Understanding lstm networks (August 2015), http://colah.github.io/posts/2015-08-Understanding-LSTMs/, accessed on March 10, 2021
13. Oliveira, J.N.: Hitex: Um sistema em desenvolvimento para historiadores e arquivistas. Forum (1992)
14. OpenNLP, A.: Welcome to apache opennlp (2017), https://opennlp.apache.org/, accessed in 18-10-2020
15. Pires, A.R.O.: Named entity extraction from Portuguese web text. Master's thesis, Faculdade de Engenharia da Universidade do Porto (2017)
16. Rodrigues, A.M., Guimarães, C., Barbedo, F., Santos, G., Runa, L., Penteado, P.: Orientações para a descrição arquivística (May 2011), https://act.fct.pt/acervodocumental/documentos-tecnicos-e-normativos/
17. spaCy: spacy 101: Everything you need to know · spacy usage documentation, https://spacy.io/usage/spacy-101, accessed in 07-01-2021
18. spaCy: Model architecture (2017), https://spacy.io/models, accessed in 14-01-2021
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I.: Attention is all you need. vol. 2017-December (2017)