Named Entity Recognition Applied to Portuguese Texts from the XVIII Century

Leonardo Zilio[1][0000−0002−6101−0814]*, Maria José Bocorny Finatto[2][0000−0002−6022−8408], and Renata Vieira[3][0000−0003−2449−5477]

[1] University of Surrey, UK, l.zilio@surrey.ac.uk
[2] Federal University of Rio Grande do Sul, Brazil, mariafinatto@gmail.com
[3] CIDEHUS, Universidade de Évora, Portugal, renatav@uevora.pt

Abstract. Extracting data and knowledge dispersed along Portuguese old medical records is important especially for researchers dealing with historical epidemiology and health sciences. An essential task in Natural Language Processing for processing textual information is Named Entity Recognition (NER). In this paper, our main objective is to test the performance of NER systems for Portuguese for extracting information from XVIII-century medical texts, so that we can provide an annotated version of an important work of this type.

Keywords: NER · XVIII-Century Portuguese · Historical Medicine.

1 Introduction

Besides the new advances in Deep Learning, Machine Learning and Neural Models, text mining techniques offer an important input for extracting knowledge from large collections. These techniques can help the work of philology researchers, historians and physicians that deal with the history of Medicine and Epidemiology. They provide effective ways to obtain, integrate and interpret data collected from different sources, and they can reduce the required time to process information from vast textual material in a non-linear approach [3].

Extracting data and knowledge dispersed among old medical records is important especially for researchers dealing with the historical epidemiology (HE) and health sciences [16]. HE holds the promise of creating a more robust and nuanced foundation for global public health decision-making by deepening the empirical records from which we draw lessons about past interventions. Several of these interventions are narrated, for example, in medical manuals published in Portuguese in the XVIII century. However, facing the complexity of old texts and understanding the information they bring is not a trivial task.

Within the scope of Information Science and the data systematization of dispersed texts [10], the gap in working with old documents and collections is also recognized as a challenge. Although there is great interest in the consideration of

temporal aspects associated with Information Retrieval (IR), Schiel [10] states that he had not found any work focused on determining the temporal context of a concept and its correlations according to the system of the time involved or even relating it to the frameworks of current knowledge.

To process and represent this kind of textual information, an essential task in Natural Language Processing (NLP) is Named Entity Recognition (NER). It corresponds to the recognition and categorization of entities mentioned in a text sample or corpus. Examples of named entities are proper names, events, places, temporal and quantitative data, etc. These extracted entities can then be further mapped to a knowledge base (KB), in special a Digital Humanities KB (DHKB) [2]. These DHKBs allow for the identification and combination of existing knowledge about historical facts from different sources.

In this paper, our main objective is to identify a NER system for Portuguese that would work for extracting information from XVIII-century medical texts[4]. To this end, we first create a gold standard based on modernized transcriptions of three text samples and then evaluate the systems' named entity (NE) extraction against this gold standard. As a second step, we contrast the systems' NE extraction performed on the modernized transcriptions against their own NE extraction from non-modernized transcriptions, which present different spelling and syntax. After evaluating the best NER system based on these two experiments, our next objective is to conduct a full NE extraction from an XVIII-century medical corpus.

The remainder of this paper is divided as follows: Section 2 presents previous studies related to NER and old texts; Section 3 describes the corpora, the gold standard, and the three NER systems that we used, while also explaining the procedures for two experiments; Section 4 contains the results of the experiments, presenting a quantitative and qualitative analysis; finally, in Section 5, we recap the main contributions of this paper and discuss future research.

2 Related Work

Quaresma & Finatto [7] developed a set of initial experiments involving NER using the medical Spanish handbook *Observaciones de Curvo*, written by Francisco Suarez de Ribera in 1735, which is based on the 1707 Portuguese Curvo Semedo's work *Observaçoes medicas doutrinaes de cem casos gravissimos*. The processing steps were done by applying NLP tools without any human intervention: from the OCR output to the creation of an ontology. Considering a sample of 10% of the extracted NEs, the authors identified a precision of 21% for locations, 22% for persons, and 5% for events. The authors also report that most of the errors occurred because of the low quality of the OCR result.

Regarding previous work on NER applied to old Portuguese, we have the work done with the Parish Memories. A digital version of these manually tran-

[4] An ongoing corpus composed of medical manuals is available on the Historical Terminology section of the Textecc project `http://www.ufrgs.br/textecc/terminologia/`.

scribed texts is freely available through the CIDEHUS Digital Portal[5]. From this collection a named entity dataset was automatically built using machine learning and language models [9]. The initial entity categories considered were person, location, and organization. This resource was made available to the community, where texts are given with their respective lists of named entities [15]. It was based on a completely automated process, with reported accuracy measures about 50%, where no new training was performed.

There are studies involving NER in other historical languages. Hubková et al. [4] present a study for NER in a Czech historical corpus. They developed a new annotated dataset for historical NER, composed of historical newspapers, and conducted experiments using recurrent neural networks, achieving performance around 70%. For medieval French, Aguilar and Stutzmann [1] present a corpus and trained a new system for legal documents of the XIII and XIV centuries. Their performance measures are around 90%.

3 Methodology

In this section, we describe our corpus. We then go through the process of creating a gold standard from a manual annotation of NEs, we describe the process of using off-the-shelf NLP tools to recognize NEs, and finally we explain our evaluation approach.

3.1 Corpus

In order to test NER in a broader set of historical texts, we collected three Portuguese text samples from the same period, written by different authors, from different domains, and using a different writing style.

The first sample is from a medical handbook written by João Curvo Semedo (1635-1719), a Portuguese physician from Monforte, in 1707: *Observaçoens medicas doutrinaes de cem casos gravissimos* [11]. This handbook presents observations made by Semedo related to diagnosis and treatment of a hundred severe cases. It contains rich information about the medical terminology existing at that time, including names for treatments and diseases. Here is an extract from Semedo's handbook which preserves the original spelling:

> **OBSERVAÇAM XLV.**
>
> De hum mercador , a quem repentinamente assaltou huma dor de colica taõ intoleravel , que estando na fé sacramental para commungar , o naõ pode fazer ; e sendo eu chamado , conheci dos grandissimos ardores ; e continuos desejos de ourinar , e vomitar , das picadas da bexiga , e do adormecimento da perna direita , que a tal dor era nephritica ; para cujo remedio appliquei hum vomitorio de tres onças de agua benedicta vigorada , e tres ajudas feitas de cozimento de rim de vacca , [...]

[5] http://www.cidehusdigital.uevora.pt

The second text sample come from the *Gazetas Manuscritas da Biblioteca Pública de Évora*[6] [5] and from the *Sermons* written by Fr. Vieira. The Gazetas is a large corpus of journalistic texts from the XVIII century[7]. The excerpt below contains a transcription that used modernized spelling:

> **Diário de 23 de agosto de 1729**
>
> Pelas cartas de Vasco Fernandes César, se soube a notícia, que aqui todos ignoravam, de que El-rei o tinhafeito Conde de Sabugosa vila junto a Viseu de que não sabemos se lhe desse senhorio.
> Chegou Rodrigo César, gordo, mas não cheio, mostrou grande desinteresse; as minas que descobriu temgrande quantidade de ouro, e se achou um grão de meia arroba, porém é mau o clima, [...]

Fr. Antônio Vieira (1608-1697) was born in Portugal, and his works are known to this day for the complexity of the text and the sophistication of the argumentation. One collection of his sermons is partially available — in a transcribed version — at the Tycho Brahe Corpus [12]. Here is an excerpt from the beginning of the sermon that we used in this study, in a partially modernized spelling:

> **SERMÃO | da | Primeira Dominga do Advento**
>
> Prégado na Capella Real, no Anno de 1652 | Amen dico vobis, non praeteribit | generatio haec donec omnia fiant. | Lucas, XXI | I
> Muitas coisas sabemos deste grande dia, todas grandes e temerosas, e duas só ignoramos. Sabemos que antes do dia do Juízo , o sol, que soía fazer o dia, se há-de escurecer e esconder totalmente com o mais horrendo e assombroso eclipse que nunca viram os mortaes. [...]

These three text samples have approximately the same size (around 2000 words). We used both modernized and non-modernized transcriptions as basis for this study.

We also have 90 transcribed observations available out of the 101 total observations present in Semedo's medical handbook, and we use this corpus for extracting NEs at the end of this study.

3.2 Gold standard

The first step for testing how NER tools perform in historical texts was to create a gold standard using our existing samples. The gold standard was generated using the modernized versions of our three text samples. Two linguists, authors of this paper, annotated all samples independently and exhaustively.

[6] From now on just referred to as the Gazetas, for short.
[7] The extract used in this study comes from the period between 1729 and 1731.

After the first round of annotation, we compared both lists and agreed upon the entries that would go into the gold standard. This final list has a total of 262 NEs. Table 1 shows the tags in the gold standard, along with a few examples[8].

Tag	Freq	Examples
PERSON	163	Conde de Sarzedas, Deus, Antonio Simões Lopo
LOCATION	44	Lisboa, Espanha, Igreja de Santo Antônio da Sé
ORGANIZATION	5	Conselho da Fazenda, Santo Ofício, doze Apóstolos
TIME	16	23 de agosto de 1729, Primavera, princípios de outubro
WORK	9	Diário de Lisboa, Polyanthea, Sagradas Escrituras
EVENT	10	Dia do Juízo, Encarnação do Verbo
VALUE	4	8.000 cruzados, dois mil cruzados
OTHER	11	arábios, Gentios, Providência Divina

Table 1. NE tags frequency in the corpus and examples.

3.3 Off-the-shelf NER tools

In this section, we briefly describe the three NER models that we used. Two pre-trained models were taken from spaCy's[9] NER libraries, and the third one is a BERT-CRF model trained for Portuguese.

The pre-trained NER models for Portuguese offered by spaCy come with different language model (LM) sizes. For this study, we selected both the large and small LMs (spaCy_lg and spaCy_sm, respectively). These models were trained on the WikiNER annotation [6], which do not contain all tags present in our gold standard[10]. When contrasting the annotations of the spaCy models with our gold standard, we considered that EVENT, TIME, WORK, and VALUE instances recognized as MISCELLANEOUS were correct.

Souza et al. [13] trained a BERT-CRF model (i.e. a BERT-based embedding model associated with a Conditional Random Fields layer) based on BERTimbau [14], which is a BERT-based embeddings model for Portuguese. BERT-CRF used the HAREM corpus [8] for NER training. HAREM contains a series of NE tags: PERSON, LOCATION, ORGANIZATION, TIME, VALUE, ABSTRACTION, EVENT, THING, WORK, OTHER. When contrasting our gold standard with the extraction made with BERT-CRF, we evaluated ABSTRACTIONS and THINGS to OTHER.

3.4 Evaluation approach

For the evaluation shown in the next section, we used a glossary of the NEs that considers only the surface form and frequency in each of the text samples. This

[8] The complete list contains additional annotations with tags that were not taken into consideration in this study (*e.g., MEASUREMENT, SYMPTOM*. The complete annotation is available at https://github.com/uebelsetzer/NER_for_Portuguese_XVIII-Century_Texts/tree/main/gold

[9] `https://spacy.io/api/entityrecognizer`

[10] SpaCy uses PERSON, LOCATION, ORGANIZATION and MISCELLANEOUS.

means that we analyzed whether the systems were able to extract the correct NEs from the samples, without taking the source segments into consideration.

We first compared the systems' performance on the modernized versions of the samples against the gold standard. Then, as a second verification, we made an intra-system comparison between the extraction from the modernized and the non-modernized samples (see Section 3.1). Finally, we selected the best system to extract NEs from the larger corpus of Semedo's work.

4 Results

In this section we present the results of a quantitative and qualitative evaluation of the named entity recognition (NER) performed by the tools. At the end we describe the annotation of NEs based on the main corpus.

4.1 NER on modernized samples

Table 2 shows an overview of the results for each NER system. It presents details of the number of partial recognition (*i.e.*, the NE tag was correct, but either only part of the NE was extracted by the system or more tokens were considered as part of the NE — *e.g.*, "Mahé do Samorim" instead of "praça Mahé do Samorim", or "Casa Antônio de Saldanha" instead of "Antônio de Saldanha") and the number of wrong tags annotated in the corpus (*i.e.*, the NE was correct, but the tag was wrong — *e.g.*, "Conde de Avintes" [Count of Avintes] was annotated as LOCATION). There is also a category for NEs that were both only partially recognized and had an incorrect tag. Finally, there are missing instances (*i.e.*, for when the system failed to identify an NE) and instances that were wrongly extracted (*i.e.*, annotated instances that were not NEs).

	Correct	**Missing**	**Pt***	**WT**	**Pt and WT***	**Wrong NER**
Bert-CRF	194	14	9	30	3	18
spaCy_lg	166	27	16	42	6	25
spaCy_sm	139	23	15	62	5	35

* Some NEs from the gold standard were extracted as in two parts, so there might be extra instances in these categories.

Table 2. Results from the 3 systems contrasted against our gold standard. (Pt: partial match; WT: wrong tag.)

Looking at the annotations generated by each system, it becomes clear that BERT-CRF had a better precision for the annotation of NE tags. Some recurrent mistakes of this system were the annotation of locations as persons and also the annotation of any digit as a value. It also missed some values when the number was written in its expanded form (*e.g.*, "150.000 réis" was correctly recognized, but "um conto de réis" [1.000 réis] was not). Regarding both spaCy models, the

main errors concentrate in the incorrect annotation of tokens as NE, and also in the annotation of persons as locations (the opposite of BERT-CRF model).

When looking at the annotations across the three text samples, it became clear that the sermon by Fr. Vieira was the most complex for the spaCy systems to properly tag the instances, as many persons were wrongly tagged (16 for spaCy_lg and 20 for spaCy_sm out of 39). However, it was Semedo's text the most complex in terms of proper recognition. The whole text had only 14 NEs in the gold standard, and BERT-CRF incorrectly identified 6 extra NEs, while spaCy_lg got 14 extra NEs, and spaCy_sm got 18 extra NEs, which is more than the number of correct NEs existing in the text. In the Gazetas, which had a total of 171 NEs, all systems worked fairly well, but BERT-CRF again showed the best performance, with 143 fully correct NEs, 14 wrong tags, and only 4 missing NEs. The spaCy_lg model had 127 correctly annotated NEs, 17 wrong tags, and 14 missing annotations; while spaCy_sm had the worst result, with 105 correct annotations and 42 wrong tags.

Another error that was common for the spaCy models (especially spaCy_sm) was the tagging of extra tokens preceding or following a NE. When there was a capitalized word in the proximity, it was often considered as part of the NE, leading to the extraction of NEs such as "Batizou Francisco de Almada" [Baptized Francisco de Almada] and "O Marquês de Marialva" [The Marquess of Marialva][11]. This was not a problem at all for BERT-CRF.

BERT-CRF had the best result when analyzing the extraction from a modernized version of the texts. However, since the process of modernizing these texts is similar to the process of a translation, it is unrealistic to expect all historical texts to be translated before applying a NER system to them. So we cannot use an NE extraction based on modernized versions as a parameter for old texts. As such, we still had to see how NER systems would work in a non-modernized version. This is what we explore in the next section.

4.2 NER systems: modernized vs. non-modernized extraction

In this section, we describe the results of our intra-system comparison. This was done to check how the results differed when the sample text varied, and the non-modernized transcription was used for the extraction of NEs.

	Full match	**Different tag**	**Partial match**	**Extra instances**	**Missing instances**
BERT-CRF	221 (83.71%)	24 (9.09%)	13 (4.92%)	6 (2.27%)	14
spaCy_lg	189 (53.54%)	32 (9.07%)	24 (6.80%)	108 (30.59%)	15
spaCy_sm	195 (63.73%)	45 (14.71%)	18 (5.88%)	48 (15.69%)	17

Table 3. NER in modernized VS. non-modernized texts.

Table 3 shows how the systems compare among themselves when ran on modernized and on non-modernized versions of the same text. The percentages

[11] In the evaluation, these were considered as partial annotations.

in brackets are a direct comparison between the annotated NEs in the non-modernized versions against the ones annotated in the modernized version (e.g. the 221 full-match instances were present in both BERT-CRF annotations with the same tag — albeit with different spellings — and they account for 83.71% of the annotations in the non modernized version). The missing annotation are the NEs in the modernized version that where not annotated in the non-modernized version. While here there was no judgment in terms of the extracted NEs themselves being correct or not, it was possible to see some interesting annotations. For instance, in the results from BERT-CRF, where the annotation of the non-modernized version came up with a few more accurate results (*e.g.* "senhora Condessa da Atalaja D . Francisca" was a partial, and "senhora Condessa de Arcos" was not present in the annotation of the modernized text). For the spaCy models, however, we see that too many extra instances were added to the annotation, and many of these extra instances were wrong, while there was a larger number of missing instances.

4.3 Annotation of Semedo's Work

From a qualitative point of view, the annotation on the non-modernized transcriptions was not as good as in the modernized versions, which was expected. In the non-modernized versions, there are spelling issues that introduce noise in the results. However, even in such adverse context, BERT-CRF annotations were still consistent, and it proved to be the most robust of the three models, as it was able to handle one of the main issues of working with old texts: the non-standard spelling.

Following our objective of retrieving information from historical medical texts, we automatically annotated NEs in Semedo's *Observaçoens medicas doutrinaes de cem casos gravissimos* [11]. The annotation contains the following distribution of tags (number of unique forms in brackets): ABSTRACTION: 335 (135); EVENT: 1 (1); LOCATION: 291 (184); ORGANIZATION: 31 (26); OTHER: 18 (13); PERSON 1326 (692); THING: 110 (54); TIME: 71 (62); VALUE: 293 (70); WORK: 30 (24); total: 2506 (1261)[12].

5 Final Remarks

We compared the performance of three off-the-shelf NER systems in Portuguese texts written in the XVII and XVIII centuries. We collected three modernized text samples from that period, annotated them with NE tags to create a gold standard, and evaluated the extractions of the three systems against this gold standard. We also compared the extractions from the modernized versions of the samples against their non-modernized versions to see how much the differences in spelling and formatting would interfere with the systems' performance.

[12] This annotation is readily available at `https://github.com/uebelsetzer/NER_for_ Portuguese_XVIII-Century_Texts` .

After analyzing the results of both experiments, we concluded that BERT-CRF had better performance, even when considering the original spelling of the historical texts. Both spaCy models had issues in recognizing NEs, changing the tag of many entities and adding wrong NEs, especially in the non-modernized versions of the texts. Considering these results, we used BERT-CRF to annotate a large sample of non-modernized texts extracted from Semedo's work *Observaçoens medicas doutrinaes de cem casos gravissimos*.

In future research, we intend to annotate this corpus with other types of entities that are relevant for understanding the medical practices of the XVII-XVIII century. Our plan is to then train a system to identify these entities and extract more information from similar texts from that period.

Acknowledgments

The authors would like to thank the following institutions for providing funding for this research: Expanding Excellence in England (E3) Fund, promoted by Research England; CNPq and FAPERGS - Brazil (FAPERGS - CAPES - 06/2018 - internacionalização - proc. 19/2551-0000718-3; CNPq 06/2019 - Productivity in research – proc. 308926/2019-6); and the Portuguese Foundation for Science and Technology (FCT), projects CEECIND/01997/2017 and UIDB/00057/2020.

References

1. Aguilar, S.T., Stutzmann, D.: Named entity recognition for french medieval charters. In: Workshop on Natural Language Processing for Digital Humanities (2021)
2. Golub, K., Liu, Y.H.: Information and knowledge organisation in digital humanities: Global perspectives (2022)
3. Higuchi, S., Freitas, C., Cuconato, B., Rademaker, A.: Text mining for history: first steps on building a large dataset. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
4. Hubková, H., Kral, P., Pettersson, E.: Czech historical named entity corpus v 1.0. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4458–4465. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.549
5. Lisboa, J.L., dos Reis Miranda, T.C., Olival, F.: Gazetas Manuscritas da Biblioteca Pública de Évora. Vol. 1 (1729-1731). Publicações do Cidehus (2018)
6. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from wikipedia. Artificial Intelligence **194**, 151–175 (2013)
7. Quaresma, P., Finatto, M.J.B.: Information extraction from historical texts: a case study. In: DHandNLP@ PROPOR. pp. 49–56 (2020)
8. Santos, D., Seco, N., Cardoso, N., Vilela, R.: Harem: An advanced ner evaluation contest for portuguese. In: quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006) (2006)

9. Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., Vieira, R.: Assessing the impact of contextual embeddings for portuguese named entity recognition. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 437–442. IEEE (2019)
10. Schiel, U.: Texto & contexto: por uma recuperação da informação com mais semântica. Ciência da Informação **50**(2) (2021)
11. Semmedo, J.C.: Observaçoens medicas doutrinaes de cem casos gravissimos (1707)
12. de Sousa, M.C.P.: O corpus tycho brahe: contribuições para as humanidades digitais no brasil. Filologia e Linguística Portuguesa **16**(esp.), 53–93 (2014)
13. Souza, F., Nogueira, R., Lotufo, R.: Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649 (2019), `http://arxiv.org/abs/1909.10649`
14. Souza, F., Nogueira, R., Lotufo, R.: 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)
15. Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., Santos, I.: Enriching the 1758 portuguese parish memories (alentejo) with named entities. Journal of Open Humanities Data **7**, 20 (2021)
16. Webb, J.: Historical epidemiology and global health history. História, Ciências, Saúde-Manguinhos **27**, 13–28 (2020)