ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2025-01-13

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Bico, M. I., Baptista, J., Batista, F. & Cardeira, E. (2024). Recognizing and linking named entities in Portuguese medieval texts. In Sílvia Araújo; Micaela Aguiar; Liana Ermakova (Ed.), Digital humanities looking at the world: Exploring innovative approaches and contributions to society. (pp. 329-337). Cham: Springer Nature.

Further information on publisher's website:
10.1007/978-3-031-48941-9_25

Publisher's copyright statement:
This is the peer reviewed version of the following article: Bico, M. I., Baptista, J., Batista, F. & Cardeira, E. (2024). Recognizing and linking named entities in Portuguese medieval texts. In Sílvia Araújo; Micaela Aguiar; Liana Ermakova (Ed.), Digital humanities looking at the world: Exploring innovative approaches and contributions to society. (pp. 329-337). Cham: Springer Nature., which has been published in final form at https://dx.doi.org/10.1007/978-3-031-48941-9_25. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

Chapter 25: Recognising and Linking Named Entities in Portuguese Medieval Texts

Maria Inês Bico[1,2][0000-0002-6280-9417], **Jorge Baptista**[3,4][0000-0003-4603-4364], **Fernando Batista**[4,5][0000-0002-1075-0177], **Esperança Cardeira**[1,2][0000-0003-4700-9830]

[1] Univ. Lisboa - Fac. Letras, Alameda da Universidade, Lisboa, Portugal
[2] Centro de Linguística da Universidade de Lisboa, Portugal
[3] Univ. Algarve - Fac. Ciências Humanas e Sociais, Campus de Gambelas, Faro, Portugal
[4] INESC-ID Lisboa - Human Language Technology Lab, R. Alves Redol 9, Lisboa, Portugal
[5] ISCTE - Instituto Universitário de Lisboa, Av. Forças Armadas, Lisboa, Portugal
mariainesb1@campus.ul.pt, jbaptis@ualg.pt, fmmb@iscte-iul.pt, ecardeira@campus.ul.pt

Abstract
Despite the continuous development of approaches and tools for Named Entity Recognition (NER), historical texts still face issues that modern ones do not have. These issues relate to the nature and type of the documents, the time period when they were produced, the diachronic difference in the language and how they were extracted from their source and preserved in their digital form. This paper addresses the challenges of identifying, recognizing, and categorizing named entities (NE) in Old and Middle Portuguese. We also briefly address the challenge of disambiguating Named Entities. The Portuguese Corpus of Ancient Texts consists of texts dating back to the 13th century up to 1525. All texts are transcribed and preserved with little editorial intervention, in XML format, using the web-based platform TEITOK. A part-of-speech (PoS) automatic annotation model was created and applied to six texts of the corpus, to enrich them and improve search queries. Following the automatic annotation task, a manual correction step ensued, so that more than half a million tokens have been lemmatized and annotated with their respective part-of-speech and inflection. For the NER task, the method's pipeline is presented, following the annotation of NE of a corpus of +400k tokens. In addition, the criteria for determining the boundaries of the NE were established, namely for the identification of name-chains and composite entities that comprise more than one word. For this task, the immediate context of the words was considered. Concerning disambiguation, this paper presents how ambiguity is pervasive within the texts and across the corpus.

1. Introduction

In corpora of historical documents, the recognition and classification of Named Entities (NE) can help in the identification of the *auctoritates* in each text and time period, or the study of the evolution of names or forms of address. This should be completed by entity linking, that is, by giving each entity an unique code that redirects to an entity entry in a knowledge base (for instance, Wikidata). Named Entity Recognition (NER) is the Natural Language Processing (NLP) task of extracting Named Entities (NE) from unstructured texts [1]. These NEs are linguistic expressions that refer to people, place names, organizations, dates, and other numeric expressions or units of information. Whichever the domain, the identification of NEs is an important task to better understand the content of any text. The use of NER in the

Humanities is becoming widespread as more corpora are made available online [2]. However, for historical documents, NER tools and algorithms still face challenges related to the nature of these documents.

This paper addresses the NER task challenges in the *Corpus de Textos Antigos* (henceforward, CTA), which deals with editions of texts in Old and Middle Portuguese. We also show how significant the lack of proper resources is and the efforts made in order to create a NE annotated corpus for these periods of the Portuguese language. The issue of NE ambiguity and the challenge of disambiguation are also briefly addressed.

Though it is nowadays possible to use off-the-shelf NER tools, they do not have the desired performance on historical documents. This is in part due to the challenges these documents pose for NLP methods, tools, and linguistic resources. If differences in the domain have an impact on modern language NER tools, it is likely that the same is also true for historical varieties of languages and various source documents. Thus, a set of challenges arise from historical documents. In the absence of an orthographic standard, multiple graphic forms exist for the same word. This graphic variation is both expected and troublesome for NER tools. Also, the level of input noise by automatic text acquisition methods (for instance, optical character recognition from facsimiles) as well as other external factors [3-4] impacts the performance of NER tools. Finally, the lack of adequate resources (for instance, corpora and lexicons) conditions all other challenges.

2. Related Work

The HAREM initiative offers a golden standard for contemporary Portuguese NER [5-6] and the *Corpus Informatizado do Português Medieval* (CPIM) [7] and *Tycho Brahe—Parsed Corpus of Historical Portuguese* [8] are corpora for historical Portuguese.

[9] enriches a corpus of the 1758 Parish Memories from the Portuguese region of Alentejo with person (PER), location (LOC) and organization (ORG) annotations. Two off-the-shelf NER tools for contemporary Portuguese were chosen: spaCy and BCF (BiLSTM-CRF+FlairBBP), and a partial manual annotation task was done following the guidelines set by the HAREM project. BCF, though more conservative, achieved better results (precision = 52.84%) and was chosen for the annotation task. Lemmatization and PoS-tagging were possible solutions thought to subdue the challenges posed by graphic variation.

[10] applied spaCy and BERT-CRF to medical texts by João Curvo Semedo (1635-1719). The two systems were applied to modernized and non-modernised transcriptions of the texts and the results were compared. Evaluation was done considering partial recognition. BERT-CRF had a better performance than spaCy, and both models had a lower performance on the non-modernised versions. BERT-CRF was able to overcome some of the spelling issues and, for this reason, it was used to annotate a larger sample of Semedo's work.

For custom NER models, [11] developed a system for the identification and categorization of entities from a medieval Spanish corpus (12th - 15th century). This system, which proposes a complex TEI-based taxonomy with nested entities, is based on entity detection rules and lexical resources available for medieval Spanish. The results were positive, especially considering the spelling variability (overall F1= 0.77%). Custom rule-sets to overcome the evolution of the language proved successful.

[12] used a Conditional Random Fields model (CRF) for the recognition and classification of entities in a corpus of medieval Latin charters from the region of Burgundy (9th - 14th century). This model takes into account different types of information at a token

level, such as lemma, PoS, case and suffix. The overall result of the general model was satisfactory (F1=0.91% and 0.95% for location and person, respectively). It was concluded that the token-level information improves the performance of the model in the recognition of some compound names, though more complex and nested entities still remain a challenge.

3. Corpus de Textos Antigos

The Corpus de Textos Antigos (CTA) (http://teitok.clul.ul.pt/cta/ accessed: 2023/03/28) was created in 2015 by CLUL's Philology group. This project aims to publish hagiographic, spiritual and didactic texts, written in or translated to Portuguese, up to 1525 and with high fidelity to the source material. The corpus was built within the TEITOK framework [13-14], and it has, as of March 2023, 32 diplomatic editions of 27 texts, from different dates and length. An effort was made in order to supply the corpus with morphosyntactic information [15], ending with 6 texts and approximately half a million tokens annotated for part-of-speech, lemma and grammatical values (inflexion). An automatic annotation task was performed using TreeTagger [16-17] and precision varied from 67.4 to 91%, depending on different factors, such as the text and the number of previously unseen tokens. These results, though satisfactory, can still be improved, namely by expanding the training dataset.

4. NER in Corpus de Textos Antigos

To this end, we set out to create a corpus for Old and Middle Portuguese annotated for NE. Two out of the six morphosyntactically annotated texts were selected for additional annotation. To ensure a large coverage of NE, the two largest texts were selected: ms. A of Horto do Esposo (henceforward, HdE-A), with +150k tokens, and Livro dos Mártires (henceforward LdM), with +250k tokens. The BIO annotation scheme [18] was adopted. The annotation task consisted in classifying the names of people (PERSON) and place names (PLACE). The subtype PERSON:ORG was defined to include organizations, such as *a Igreja* 'the Church', or collective bodies of people, such as *Israel* when it is used to identify the Jewish people.

The guidelines for NE boundaries were adopted from the HAREM project [5-6]. The principle that all NE must have a proper noun was kept, though some exceptions were allowed for entities such as *Deus* 'God' or *Igreja* 'Church', which had been annotated as common nouns. Entities have a flat representation (no nesting is allowed). All appositions, epithets and forms of address, as well as punctuation tokens, were included, so that each NE is delimited at its maximal extension (for instance, *Maximo o notario do Emperador* 'Maximo, the Emperor's notary'). The rationale for this option is that the incorporation of contextual information could be helpful later on to disambiguate entities (for instance, *Carlos, rei de França* 'Charles, king of France' and *Carlos, rei de Inglaterra* 'Charles, king of England'). An example of the BIO-annotation for the PERSON and PLACE types is shown on Table 25.1. Table 25.2 shows the breakdown of entities' types per text. Most entities are composed of only one or two tokens (92.74% and 71.24% for HdE-A and LdM, respectively). The average length of the NEs is 1.6 tokens for HdE-A and 2.3 tokens for LdM. The slightly larger extension of entities in LdM can be explained by the text's long descriptive appositions and epithets.

Some found patterns can already be used for NER. A recurring two-token pattern for the PERSON entity is the combination of the adjective *são/santo/santa* 'saint' followed by a

proper noun: *são Paulo*, *santo António*, *santa Justina*; as well as the titles for emperors and prophets (*imperador Nero*, *profeta David*). For PLACE names, classifiers such as *cidade de* 'city of' or *terra de* 'land of' are very common, for instance, *cidade de Roma*, *terra de Escócia*. Some irreducible variation must, however, be taken into consideration: for instance, *a cidade celestial de Jerusalém* 'the celestial city of Jerusalem', where the adjective is not a facultative but an essential element for the delimitation of the NE; or in *bem-aventurado mártir são Zoil* 'blessed martyr saint Zoil', where the classifier *mártir* is freely modified by the adjective *bem-aventurado*.

Table 25.1

Example of BIO annotation for PERSON and PLACE

Word	nosso	senhor	jhesu	cristo	|	cidade	de	Roma
Lemma	meu	senhor	Jesus	Cristo	|	cidade	de	Roma
	my 1st pl	*lord*	*lord*	*Christ*	|	*city*	*of*	*Rome*
POS	PRO+pos:1pms	N:ms	N+Npr:ms	N+Npr:ms	|	N:fs	PREP	N+Npr:fs
BIO tag	B	I	I	I	|	B	I	I
Type	PERSON	PERSON	PERSON	PERSON	|	PLACE	PLACE	PLACE

Table 25.2

Total of annotated NE in HdE-A and LdM

NE Type	PERSON	PLACE	PERSON:ORG	Total
HdE-A	2,785 (84.44%)	319 (10.13%)	45 (1.43%)	3,149
LdM	8,507 (90.11%)	920 (9.74%)	14 (0.15%)	9,441
Combined Total	11,292 (86.69%)	1,239 (9.84%)	59 (0.47%)	12,590

Automatically linking a NE to an unique ID in a knowledge base is a challenging task. Naturally, some entities are clearly unambiguous in these texts (for instance, *Beda* 'Bede', *Aristoteles* 'Aristotle', *Eva* 'Eve'). However, the graphical variation found in these documents (for instance, *Simão* 'Simon' and its variants: *Simaã*, *Symon*, *Symõ*, *Simõ*, *Symão*, *Symaon*); or the name's commonness, for instance, *João* 'John', or its regular use in certain contexts, for instance, *Papa Clemente* 'Pope Clement' (which one of the 14 popes so named?); hinders the automatic recognition of unique entities. It seems that, alongside an automatic approach, the use of contextual information and manual verification is needed.

5. Final considerations

This chapter presented the groundwork for the NER task on the *Corpus de Textos Antigos* and its main challenges. This groundwork consists of the creation of a corpus for Old

and Middle Portuguese annotated for its named entities. This corpus is made of two different texts, already previously annotated with the words' lemma and PoS-tags, with approximately 400 thousand tokens. We have considered Named Entities of the type: PERSON (for instance, *Santo Agostinho* 'Saint Augustine'), PERSON:ORG (*Santa Madre Igreja* 'Holy Mother Church') and PLACE (*Lisboa* 'Lisbon'), which were annotated with the BIO formalism, and a set of guidelines was produced for determining their boundaries. As for other historical documents, the main challenges are the multiple graphical forms a word might have within the same text and across the corpus, the irregular use of punctuation and capital letters for proper nouns, and the ambiguity of the entities. The first set of challenges can be addressed with the enrichment of the training corpus. Ambiguity is not so easily solved, though, as the association between entities and unique IDs in a knowledge database implies the use of contextual information.

All in all, the creation of a specific annotated corpus for the recognition of named entities in Old and Middle Portuguese highlights the importance of adequate resources for the different NLP tasks. Even if resources thought for contemporary languages can give some results, the use of suitable resources for both language and time period, though taxing, will increase the performance of models, and save time in later stages of research. As a resource for the study of Old Portuguese, this corpus is relevant for the study of the evolution of proper nouns, and the evolution of the language itself. Further research, which is currently underway, will shed light to some of these questions, such as the modern lemmatization of older versions of a noun or the co-existence of two different forms for the same name.

6. Acknowledgements

Research for this chapter was partially funded by public funds through Fundação para a Ciência e a Tecnologia: J. Baptista and F. Batista (INESC-ID Lisboa, Proj. ref. UIDB/50021/2020), E. Cardeira (School of Arts and Humanities, Center of Linguistics, University of Lisbon, Proj. ref. UIDP/00214/2020) and M. I. Bico by Ph.D grant (proj. ref. UI/BD/152806/2022).

References

[1] Nadeau, David, and Satoshi Sekine. 2007. A survey of named entity recognition and classification. Lingvisticæ Investigationes. John Benjamins. https://doi.org/10.1075/li.30.1.03nad.

[2] Ehrmann, Maud, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic Evaluation of NER Systems on Old Newspapers. In Proc. 13th Conf. Natural Language Processing (KONVENS 2016), 97–107. Bochum, Germany: Bochumer Linguistische Arbeitsberichte.

[3] Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named Entity Recognition and Classification on Historical Documents: A Survey. https://doi.org/abs/2109.11406.

[4] Grover, Claire, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named Entity Recognition for Digitised Historical Texts. In Proc. Intl. Conf. Language Resources and Evaluation (LREC 2008), 1343–1346.

[5] Cardoso, Nuno, and Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. In Reconhecimento de entidades mencionadas

em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área, 2011–238. Linguateca.
[6] Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas, and Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. In Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, 11–31. Linguateca.
[7] Xavier, Maria Francisca. 2016. O CIPM – Corpus Informatizado do Português Medieval, fonte de um Dicionário exaustivo. In Lingüística de corpus y lingüística histórica iberorrománica, 137–156. Berlin/Boston: De Gruyter. https://doi.org/doi:10.1515/9783110462357-007.
[8] Galves, Charlotte. 2018. The Tycho Brahe Corpus of Historical Portuguese. Romance Parsed Corpora.
[9] Vieira, Renata, Fernanda Olival, Helena Freire Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. Journal of Open Humanities Data. https://doi.org/10.5334/johd.43.
[10] Zilio, Leonardo, Maria José Finatto, and Renata Vieira. 2022. Named Entity Recognition Applied to Portuguese Texts from the XVIII Century. In Proc. 2nd Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022), 1–10.
[11] Díez Platas, Ma Luisa, Salvador Ros Muñoz, Elena González-Blanco, Pablo Ruiz Fabo, and Elena Álvarez Mellado. 2021. Medieval Spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. Journal of the Association for Information Science and Technology 72: 224–238. https://doi.org/10.1002/asi.24399.
[12] Chastang, P., S. Torres Aguilar, and X. Tannier. 2021. A Named Entity Recognition Model for Medieval Latin Charters. Digital Humanities Quarterly 15:4.
[13] Janssen, Maarten. 2016. TEITOK: Text-faithful Annotated Corpora. In Proc. 10th Intl. Conf. Language Resources and Evaluation (LREC'16), 4037–4043. Portorož, Slovenia: European Language Resources Association (ELRA).
[14] Vaamonde, Gael, and Maarten Janssen. 2020. Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK. In Calen barbas, falen cartas. A escrita en galego na Idade Moderna, 271–292. Consello da Cultura Galega (Ensaio & Investigación. https://doi.org/10.17075/cbfc.2020.008.
[15] Bico, Maria Inës, Jorge Baptista, Fernando Batista, and Esperança Cardeira. 2022. Early Experiments on Automatic Annotation of Portuguese Medieval Texts. In Linking Theory and Practice of Digital Libraries, ed. Gianmaria Silvello, Oscar Corcho, Paolo Manghi, Giorgio Maria Di Nunzio, Koraljka Golub, Nicola Ferro, and Antonella Poggi, 442–449. Cham: Springer Intl. Publishing. https://doi.org/10.1007/978-3-031-16802-4_44.
[16] Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proc. Intl. Conf. New Methods in Language Processing. Manchester, UK.
[17] Schmid, H. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In Natural Language Processing Using Very Large Corpora, 13–25. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-2390-9_2.
[18] Sang, Erik F. Tjong Kim, and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. In Proc. CoNLL-2000 and LLL-2000, 127–132. Lisbon.