Too Young to NER: Improving Entity Recognition on Dutch Historical Documents

Vera Provatorova,[†,⊕] Marieke van Erp,[†] and Evangelos Kanoulas[⊕]

[†]DHLab, KNAW Humanities Cluster
Oudezijds Achterburgwal 185
1012 DK Amsterdam
The Netherlands
{vera.provatorova,marieke.van.erp}@dh.huc.knaw.nl

[⊕]University of Amsterdam
Science Park 904
1098 XH Amsterdam
The Netherlands
e.kanoulas@uva.nl

Abstract

Named entity recognition (NER) on historical texts is beneficial for the field of digital humanities, as it allows to easily search for the names of people, places and other entities in digitised archives. While the task of historical NER in different languages has been gaining popularity in recent years, Dutch historical NER remains an underexplored topic. Using a recently released historical dataset from the Dutch Language Institute, we train three BERT-based models and analyse the errors to identify main challenges. All three models outperform a contemporary multilingual baseline by a large margin on historical test data.

Keywords: named entity recognition, digital humanities, historical texts

1. Introduction

Named Entity Recognition (NER) is the task of detecting named entities (people, locations, organisations, etc.) mentioned in text (Sang and De Meulder, 2003). NER is widely used for a range of downstream tasks in various domains, including question answering, content recommendation, conversational search and other tasks.

Making digital archives easily searchable is important for researchers in digital humanities, for example for prosopographical research (Tamper et al., 2019). A reliable NER system contributes greatly to this goal: it allows to save manual efforts in looking for information about particular people, places and other entities. However, recognising entities in historical documents is far from a straightforward task: the nature of the data leads to multiple challenges, including OCR noise, historical spelling variations, and potential differences in language use compared to modern texts. The task becomes even more challenging when the documents are written in a low- or mid-resource language: while a vast amount of training data is available for English or French, other languages are less common, leading to a relative lack of parametric knowledge.

While recent advances have been made in recognising and linking historical entities in multiple languages (Ehrmann et al., 2020, 2022), Dutch historical documents remain an underexplored domain, despite the data being publicly available (Dutch Language Institute, 2022). In this paper, we delve into Dutch historial named entity recognition; we train and test three different NER models on historical data ranging from the 17th to the 19th century and provide an extensive analysis of the performance of these models. We hope to inspire further research on Dutch historical NER and draw attention of the research community to the available language resources.

The remainder of this paper is organised as follows. In Section 2, we discuss related work in historical named entity recognition. In Section 3 we detail our experimental setup. We present our results and discussion in Section 4 and conclusions and future work are presented in Section 5. Our code is available at `https://github.com/vera-pro/Dutch-NER-LT4HALA`.

2. Related Work

Languages change over time. In particular prior to the introduction of the printing press and language standardisation language, spelling and writing style variation was widespread. Furthermore, the concepts covered in texts over longer periods of time evolve too, making the analysis and interpretation of historical texts an even greater challenge than contemporary texts (Montanelli and Periti, 2023).

Dutch is a West-Germanic language mainly spoken in the Netherlands, Belgium and Suriname. The language is similar in German in that noun compounding is productive and compounds are generally written without spaces. A term such as notarial deed, made up of 'notary' and 'akte' would thus become 'notarisakte'. The language has many loanwords from French, German and Latin. A particular peculiarity that affects named entity recognition is that it is common for family names to contain location names (Brouwer et al., 2022). Prior to the 18th century, there was no standard Dutch spelling. Although various attempts were made to establish

LT4HALA 2024@LREC-COLING-2024, pages 30–35
25 May, 2024.

dataset	century span	# entity annotations			data source
		PER	LOC	TIME	
train	17th-19th	55,921	30,636	19,809	see test: SA, test: VOC, test: RHC, test: NHA
validation	17th-19th	14,393	7,427	4,782	see test: SA, test: VOC, test: RHC, test: NHA
test: SA	17th-18th	781	257	255	Notarial deeds from the Amsterdam City Archive
test: VOC	17th-18th	290	315	180	Notarial deeds of the Dutch East India Company
test: RHC	19th	24	17	5	Notarial deeds from the archives of the Dutch regional historic centra
test: NHA	19th	352	252	109	Notarial deeds archive of Haarlem
test: CoNLL'02	21st	1098	774	0	Belgian newspaper "De Morgen" of 2000 (editions from June to September)

Table 1: Dataset details. The training and validation splits, as well as historical test splits, are part of (Dutch Language Institute, 2022). The contemporary test set is from (Tjong Kim Sang, 2002).

a guide, none gained widespread adoption. With the rise of printing, spelling standardization accelerated. Modern Dutch spelling can be traced back to the 1860s, when Matthijs de Vries and Lammert Allard proposed a set of spelling rules and word lists forming the basis of contemporary written. These efforts were supported by the government (Donaldson, 1983).[1]

Contemporary language models such as BERT (Devlin et al., 2019), Bloom (Scao et al., 2022) and LLaMA (Touvron et al., 2023) are optimised for contemporary language. This means these models may not perform as well on historical texts that differ from modern language (Hosseini et al., 2021; Lai et al., 2021). Historical texts often contain obsolete expressions or words with different meanings than today. Additionally, spelling variations and OCR errors may limit the accuracy of automated text processing systems.

The task of historical NER has been gaining popularity in the recent years, with domain-specific NER research focusing on for example medieval Latin charters (Chastang et al., 2021) or historical locations (Won et al., 2018). (Ehrmann et al., 2020) introduced HIPE, a shared task focused on recognising and linking entities in historical newspapers. Two years later, the next shared task on this topic has been introduced by the same team (Ehrmann et al., 2022). The languages in HIPE '20 include English, German and French, with Finnish and Swedish added as extra languages in HIPE '22.

The contributions most similar to ours are (Hendriks et al., 2020), where the authors performed NER and record linkage on historical Amsterdam notarial archives and personnel records of the United East Indies Company (VOC), and (Arnoult et al., 2021), where the authors experimented with Dutch and multilingual NER models on their new dataset of VOC records. As this work was done prior to the latest iteration of LLMs and the introduction of the NER dataset by the Dutch Language Institute, we further build upon and extend the understanding of NER performance on historical Dutch texts. For further reading, we refer the reader to the following historical NER surveys: (Blouin et al., 2021; Humbel et al., 2021; Ehrmann et al., 2023).

3. Experimental Setup

Following (Sang and De Meulder, 2003), we approach NER as a token classification problem. We focus on transformer-based models as these provide the best performance and ease of use in transfer learning at the time of writing (Li et al., 2020). In this section, we detail which models were used and how we fine-tuned them, the datasets we tested on, and the approach we used for evaluation and error analysis.

3.1. Models

We fine-tune three BERT-based models on historical data:

1. BERTje (De Vries et al., 2019), a Dutch model trained on a mixture of modern texts and historical novels, with modern texts being the majority in the training data;

2. GysBERT (Manjavacas and Fonteyn, 2022), a Dutch model designed specifically for historical data;

3. mBERT (Devlin et al., 2019), a multilingual model that includes Dutch as one of its languages.

The models were trained on one GPU for 15 epochs with early stopping. We used the batch size 8 and selected the best checkpoint by F1 score. To evaluate the models against a strong baseline that has not been optimised for historical data, we compare them with WikiNEuRal (Tedeschi et al., 2021). This

[1] https://www.dbnl.org/tekst/dona001dutc02_01/dona001dutc02_01_0007.php

is a multilingual NER model that includes Dutch as one of its languages and achieves high scores on contemporary benchmarks.

3.2. Datasets

We fine-tune the models using the training and validation splits of the NER dataset provided by Dutch Language Institute (2022). This dataset was created in 2020 through a crowdsourcing project initiated by the Dutch National Archive. The dataset contains notarial deeds from eleven different Dutch archives, some focused on Dutch East India Company dealings, others on local notary business. For testing the models, we use the test splits of Dutch Language Institute (2022) as well as a dataset with modern texts: the test split of Tjong Kim Sang (2002). Table 1 shows the details of the datasets. There are many different NER categorisations. In (Dutch Language Institute, 2022) the labels PER, LOC and TIME are present, while for (Tjong Kim Sang, 2002) the labels are PER, LOC, ORG, and MISC. Since the last two labels are not seen by the models in the training data, we exclude them from evaluation. As WikiNEuRal has extra NER labels in its vocabulary, we consider the predictions containing these labels as 'O' when comparing the models.

3.3. Evaluation

To identify main challenges in historical Dutch NER, we first group the data subsets by century to analyse the role of time. We analyse precision and recall of the models per century, create confusion matrices, identify overlaps in the wrong predictions made by different models, and perform qualitative analysis to find examples of challenging NER cases.

4. Results and Discussion

This section describes the results of our experiments and the error analysis. Table 2 shows precision, recall and F1 score per model per century for two NER labels, PER and LOC (TIME is excluded from this part of the analysis since WikiNEuRal does not predict it). For both labels the same pattern is observed: WikiNEuRal achieves best results on contemporary data and performs substantially worse than all other models on historical data. Interestingly, GysBERT does not outperform BERTje and mBERT on historical data, despite having seen more historical texts during pre-training: the three models achieve approximately the same results. On the contemporary test set, however, mBERT performs worse than all other models, achieving particularly low scores in both precision and recall on the LOC entity class.

Figure 1 shows confusion matrices for all labels per model per century. The main diagonal displays the number of correctly classified tokens for each label. Note that the exact number of tokens may vary per model, since each model has its own WordPiece tokenizer. From the figure we identify four most common classes of errors:

1. "False positive": predicting an entity when the correct label is "O";
2. "False negative": predicting "O" when the correct label is an entity;
3. Mention boundaries: predicting a correct class but with "I-" instead of "B-" and vice versa;
4. People vs. places: confusing "PER" and "LOC" entities.

When looking closely at the error examples during our qualitative evaluation, we noticed that some errors are caused by wrong annotations in the test sets: for example, the entity "Willem van Zonneveld" in the NHA test set is labelled as two separate PER entities, "Willem van" and "Zonneveld", which is incorrect. All models except WikiNEuRal recognise this entity correctly, which leads to a mention boundaries error. Some errors, however, are indeed caused by the models making wrong predictions: for example, in the CoNLL test set mBERT incorrectly predicts two separate LOC entities for "Los Angeles". In case of the "people vs. places" errors, qualitative analysis shows that many examples are ambiguous, and some of the mistakes made by the models could be also made by a human annotator. For example, "Jan Hendrik du Caijlar van Delf" in the VOC test set is labelled as one PER entity with a double surname, but all models predict "Delf" as a separate entity, as in "Jan Hendrik du Caijlar from Delft". This type of errors is an interesting challenge typical for Dutch texts, since Dutch family names often contain location names (Brouwer et al., 2022).

Figure 2 is a Venn diagram showing the overlap in wrong predictions between models for every test set. Note that an overlap between two models here means that both models gave a wrong answer, but the answer is not necessarily the same for the two models. The error overlap is small for all historical test sets, which indicates that the models tend to make different mistakes and therefore could benefit from ensembling.

5. Conclusion and Future Work

We used historical texts from the Dutch Language Institute to train three BERT-based NER models, making one of the first steps towards publicly available Dutch historical NER. All models are shown to

label	model	17-18 P	17-18 R	17-18 F	19 P	19 R	19 F	20 P	20 R	20 F
PER	GysBERT	.71	.67	.69	.76	**.73**	.74	.74	.76	.75
	BERTje	**.76**	**.71**	**.73**	**.80**	**.73**	**.76**	.88	.83	.85
	mBERT	.72	.68	.70	.77	.72	.74	.74	.71	.72
	WikiNEuRal	.48	.40	.43	.61	.45	.51	**.94**	**.86**	**.90**
LOC	GysBERT	.74	**.79**	.76	**.81**	**.77**	**.79**	**.72**	.66	.69
	BERTje	.77	.78	**.78**	.78	**.77**	.78	.71	.71	.71
	mBERT	**.79**	.77	**.78**	**.81**	.75	.78	.51	.48	.50
	WikiNEuRal	.48	.50	.49	.50	.48	.49	**.72**	**.90**	**.80**

Table 2: Precision, recall and F1 score per century on the PER and LOC labels.

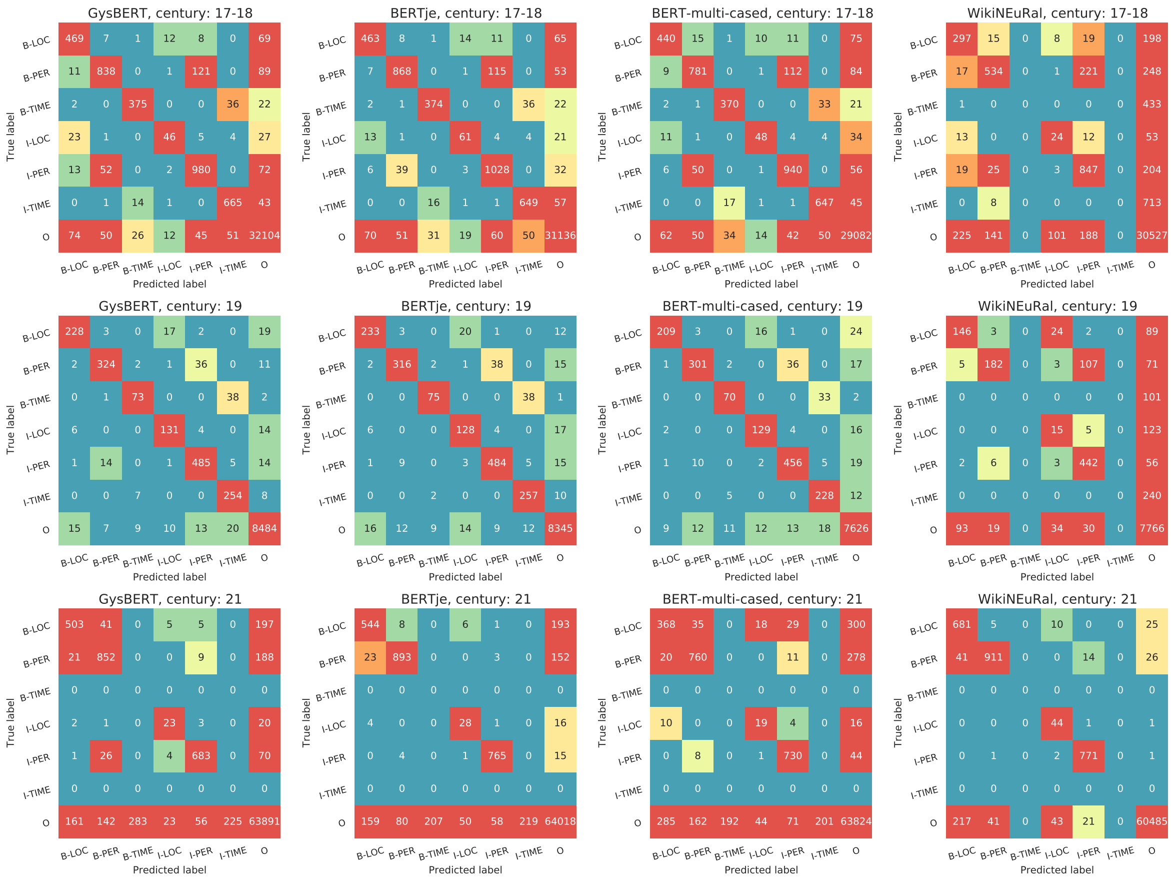

Figure 1: Confusion matrices of the models per token per century. Every cell shows a number of tokens.

perform well on historical data from the 17th to the 19th century, achieving substantially better scores than the baseline. Our error analysis shows that the overlap in wrong predictions on historical data is small, which indicates that using an ensemble of the three models might be optimal for recognising entities in Dutch historical data. Future work includes implementing and testing such an ensemble, as well as experimenting with more diverse entity types and testing on additional domains.

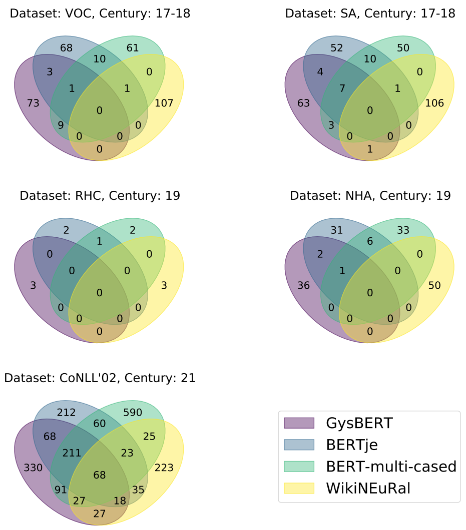

Figure 2: The overlap of false predictions per dataset. Every petal shows a number of sentences with at least one wrong prediction.

6. Acknowledgements

This research was supported by the KB National Library of the Netherlands Researcher-in-Residence program, NWO Smart Culture – Big Data / Digital Humanities (314-99-301), the Informatics Institute of the University of Amsterdam, and the European Union (grant agreement 101088548 - TRIFECTA). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We thank Sara Veldhoen, Marieke Moolenaar and Willem Jan Faber for their helpful feedback.

7. Bibliographical References

Sophie I Arnoult, Lodewijk Petram, and Piek Vossen. 2021. Batavia asked for advice. Pretrained language models for named entity recognition in historical texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 21–30.

Baptiste Blouin, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. Transferring modern named entity recognition to the historical domain: How to take the step? In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 152–162.

Leendert Brouwer, Peter McClure, and Charles Gehring. 2022. Dutch family names. In *Dictionary of American Family Names*. Oxford University Press.

Pierre Chastang, Sergio Torres Aguilar, and Xavier Tannier. 2021. A named entity recognition model for medieval latin charters. *Digital Humanities Quarterly*, 15(4).

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bruce Donaldson. 1983. *Dutch: A linguistic history of Holland and Belgium*. Uitgeverij Martinus Nijhoff, Leiden.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended overview of clef hipe 2020: named entity processing on historical newspapers. In *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*, volume 2696. CEUR-WS.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, Simon Clematide, Gulielmo Faggioli, Nicola Ferro, Alan Hanbury, and Martin Potthast. 2022. Extended overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *CEUR Workshop Proceedings*, 3180, pages 1038–1063. CEUR-WS.

Barry Hendriks, Paul Groth, and Marieke van Erp. 2020. Recognizing and linking entities in old dutch text: A case study on voc notary records. In *COLCO*, pages 25–36.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural language models for nineteenth-century english. *Journal of Open Humanities Data*.

Marco Humbel, Julianne Nyhan, Andreas Vlachidis, Kim Sloan, and Alexandra Ortolja-Baird. 2021. Named-entity recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future. *Journal of Documentation*, 77(6):1223–1247.

Viet Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event Extraction from Historical Texts: A New Dataset for Black Rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400, Online. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Enrique Manjavacas and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association (ELRA).

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Minna Tamper, Petri Leskinen, and Eero Hyvönen. 2019. Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 199–214. Springer.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. ensemble named entity recognition (ner): evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, 5:2.

8. Language Resource References

Dutch Language Institute. 2022. *AI-Trainingset for NER (Version 1.0)*. Dutch Language Institute. Dutch Language Institute, 1.0. [link].

Tjong Kim Sang, Erik F. 2002. *Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition*. [link].