Unsilencing Colonial Archives via Automated Entity Recognition

Julien Peignon
julien.peignon@ensae.fr

Abstract

This report investigates the use of neural architectures for named entity recognition on historical Dutch texts from the archives of the Dutch East India Company. Building on the annotated corpus introduced by Luthra et al. (2022), we evaluate three models: a Transformer + CRF model trained from scratch, a BERTje + CRF model leveraging pretrained Dutch language representations, and a variant with domain-adaptive pretraining. Our experiments show that the BERTje + CRF model achieves the highest performance, surpassing both the baseline reported in the original article and our untrained Transformer. These results underscore the importance of transfer learning in low-resource and domain-specific contexts such as historical document processing.

1 Introduction

1.1 Context and Motivation

Colonial archives are increasingly recognized not merely as repositories of historical information but as instruments of power that shaped, and continue to shape, historical narratives [14]. Produced largely by colonial administrators, these archives reflect the perspectives, values, and biases of those in positions of authority, often at the expense of marginalized groups such as enslaved people, women, and indigenous populations [7]. A paradigmatic example is the archive of the Dutch East India Company (VOC), which includes thousands of notarial testaments created in its overseas settlements. While these documents contain rich information about social relations in the colonial context, traditional archival tools—such as name indexes compiled in the nineteenth century—have systematically foregrounded European male actors and omitted others, thereby perpetuating archival silences [15]. In response to this exclusion, recent scholarly efforts aim to "unsilence" the archives by employing computational methods that can surface previously overlooked individuals and relationships. This shift reflects a broader commitment to decolonizing archival infrastructures and expanding access to the historical record in more inclusive and critical ways.

In response to these archival silences, the article by Luthra et al. [11] focuses on the testaments of the VOC as a case study. These documents, although rich in social detail, have long excluded non-European and unnamed individuals from formal archival access tools. By applying automated Named Entity Recognition (NER) to this corpus, the authors aim to surface marginalized voices and demonstrate how computational methods can contribute to a more inclusive reconstruction of the colonial past.

1.2 Named Entity Recognition in Historical Texts

NER is a core task in natural language processing that consists in automatically identifying and classifying mentions of entities in a text. These typically include proper nouns such as persons, organizations, and locations, but may also encompass dates, numerical quantities, or domain-specific

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

categories. In standard implementations, NER models assign one of several predefined labels to each token in a sequence, often using architectures such as Conditional Random Fields (CRF) or neural sequence models like BiLSTM-CRF and transformers [12, 9]. Accurate NER is crucial for downstream tasks including information retrieval, question answering, and knowledge base population. In archival contexts, it can facilitate access to unstructured historical records by extracting and indexing relevant entities automatically.

Applying NER to historical texts, however, poses distinct challenges. Historical corpora often contain spelling variations, archaic or regionally specific vocabulary, and noisy inputs due to Optical Character Recognition (OCR) or Handwritten Text Recognition (HTR) errors. Moreover, the semantic categories of interest in historical documents frequently diverge from standard NER schemas. For example, colonial archives may refer to enslaved individuals or indigenous groups using generic descriptors rather than proper names. Recent studies have shown that modern neural architectures, especially those leveraging contextual embeddings from pretrained language models like BERT, can be adapted to such corpora with varying degrees of success [3, 5]. The inclusion of subword embeddings and domain-adaptive pretraining have been shown to mitigate some of the noise and data scarcity issues typical of historical settings. Nonetheless, NER in this context remains a resource-intensive task that often requires task-specific annotation guidelines and typologies, as demonstrated by Luthra et al. [11].

1.3 Article Contributions

The article by Luthra et al. [11] makes several important contributions to the application of natural language processing in colonial archival research. First, it proposes a novel annotation typology specifically designed for historical documents produced in a colonial context. Unlike standard NER typologies, which typically focus on named entities such as persons, organizations, and locations, their approach explicitly includes unnamed individuals—such as enslaved persons referred to only by descriptors—as well as associated attributes like gender, legal status, and notarial role. This extension of the NER task into broader entity recognition and classification allows for a more inclusive and context-aware modeling of the archival record.

Second, the authors release a high-quality annotated corpus based on digitized VOC testaments. The dataset comprises over 68,000 annotations distributed across more than 2,000 pages, and is provided in IOB format to facilitate reuse. The annotations were created by trained annotators using carefully designed guidelines, and inter-annotator agreement metrics were reported to assess annotation consistency. To our knowledge, this is the first public corpus of its kind for Dutch colonial archives.

Third, the paper presents baseline results using state-of-the-art NER models tailored to historical Dutch-language data. The authors implement two primary modeling approaches: a traditional CRF model and a neural architecture combining BERTje—a Dutch version of the BERT language model—with a bidirectional LSTM and CRF output layer (BiLSTM-CRF). These choices are motivated by prior work showing that CRF models perform well on small, domain-specific datasets, while contextualized language models like BERTje offer strong generalization due to pretraining on large Dutch corpora. The BiLSTM layer is added to capture token-level dependencies, and the CRF layer enforces valid tag transitions in the output sequence. To evaluate model performance, the authors report precision, recall, and F1 scores using micro and macro averages. On the main entity types—*Person*, *Place*, and *Organization*—the BERTje + BiLSTM-CRF model achieves a micro-averaged F1 score of 0.63, with precision and recall of 0.71 and 0.57 respectively. In comparison, the CRF baseline obtains an F1 score of 0.63 with slightly higher precision (0.73) but lower recall (0.56), indicating its conservative prediction behavior. The BERTje model outperforms CRF on recall across all major tags. For the *Person* entity specifically, the BERTje + BiLSTM-CRF model achieves an F1 score of 0.69. However, performance on more difficult tasks—such as identifying unnamed entities or attributes like *Legal Status* and *Role*—remains modest, with F1 scores often below 0.5. These results confirm the difficulty of named entity recognition on noisy, historical texts but also demonstrate the potential of pre-trained language models when fine-tuned on task-specific annotated data.

2 Methodology

2.1 Dataset

We base our experiments on the annotated corpus released by Luthra et al. [11], which consists of notarial testaments from the VOC, dated primarily from the 18th century. These documents were originally written in early modern Dutch and concern a wide array of individuals including European colonists, local inhabitants, and enslaved persons. The corpus was digitized using the Transkribus HTR platform and post-processed to produce machine-readable text.

The released dataset comprises 2,193 unique pages, plus 307 duplicated pages used to compute inter-annotator agreement, for a total of 2,500 annotated pages. Annotations follow the Inside–Outside–Beginning (IOB) format and are provided for four entity types: *Person*, *Place*, *Organization*, and *Proper Name*. Entities of type *Person* are further enriched with three attributes: *Gender*, *Legal Status*, and *Role*. These annotations were carried out using the BRAT tool and guided by a fit-for-purpose typology that includes both named and unnamed persons, particularly to capture marginalized individuals often absent from conventional archival finding aids.

The dataset contains a total of 68,429 annotations: 32,203 entity-level spans and 36,226 attribute annotations. It is split into training (70%), validation (10%), and test (20%) sets using stratified sampling to preserve the distribution of entity types and annotators. Inter-annotator agreement, measured using Cohen's kappa, ranges from 0.5 (strict match) to 0.8 (with fuzzy offset), indicating acceptable consistency for such a complex and historically sensitive annotation task. Summary statistics on IOB tag distributions per document are provided in Table 2.

2.2 Modeling Approaches

Decoder + CRF Our first modeling approach implements a lightweight Transformer decoder architecture combined with a CRF layer for structured prediction. This model is trained from scratch, without any pretraining, to assess whether self-attention-based architectures can effectively learn to tag historical Dutch texts when provided with only a moderate amount of annotated data.

The decoder consists of a token embedding layer and a sinusoidal positional encoding module, followed by a stack of Transformer encoder layers. Each encoder layer includes a multi-head self-attention mechanism and a feed-forward subnetwork, both wrapped in residual connections and pre-layer normalization. The attention mechanism computes contextualized token representations by querying each token's relationship to others in the sequence. Specifically, queries, keys, and values are derived from linear projections of the input, and the attention weights are computed as scaled dot products between queries and keys, normalized via a softmax function [16]. These weights are then used to compute a weighted sum over the value vectors, allowing the model to dynamically attend to relevant context regardless of position.

To generate final predictions, the decoder output is passed through a linear classification head, producing token-level label logits. Instead of applying independent softmax classifiers at each position, we use a CRF layer [8]. The CRF models the entire output sequence jointly and captures dependencies between neighboring labels (e.g., ensuring that a token labeled as 'I-PERSON' follows a 'B-PERSON'). During training, the CRF maximizes the log-likelihood of the correct label sequence given the emission scores from the decoder. At inference time, it performs exact decoding using the Viterbi algorithm to return the most likely tag sequence [6].

This architecture allows the model to learn contextual representations through attention and to enforce global label consistency via the CRF head. However, as we will see in the results, the absence of pretraining limits its performance significantly on this task, likely due to the relatively small size and linguistic specificity of the annotated corpus.

BERTje + CRF Our second modeling approach leverages BERTje, a Dutch-language version of the BERT model pre-trained on a large corpus of Dutch texts, including news articles, Wikipedia, and governmental documents [2, 1]. BERTje uses the transformer encoder architecture and is trained with a Masked Language Modeling (MLM) objective to learn deep bidirectional contextual representations of tokens. This pretraining allows the model to capture rich syntactic and semantic patterns of Dutch language that are transferable to downstream tasks such as named entity recognition [13].

We combine BERTje with a CRF layer to form a robust sequence labeling architecture. Unlike our Transformer decoder trained from scratch, BERTje provides high-quality contextual embeddings from the outset, which is particularly beneficial in low-resource settings such as historical document processing. Its ability to disambiguate entity types based on broader context and to handle subword tokenization is especially advantageous when dealing with non-standardized or archaic spelling found in the VOC testaments.

Fine-tuning BERTje on our annotated dataset allows the model to adapt its representations to the historical domain, while still benefiting from general linguistic knowledge acquired during pretraining. As our results will show, this architecture yields a significant performance improvement over the untrained decoder.

DAPT + BERTje + CRF Our third approach applies Domain-Adaptive Pretraining (DAPT) to BERTje before fine-tuning it for NER. DAPT is a widely used method in transfer learning where a pretrained language model is further adapted to a specific domain using unannotated text from that domain and the same MLM objective used during its original training [4, 10]. The goal is to refine the model's internal representations so they better reflect the vocabulary, syntax, and stylistic features of the target corpus, potentially improving downstream performance in low-resource settings.

In our case, we apply DAPT using the very same corpus that we later use for fine-tuning, but stripped of its entity annotations. This choice was motivated by the fact that the VOC testaments represent a highly specialized textual domain—early modern Dutch legal prose—characterized by archaic spelling, colonial terminology, and formulaic sentence structures. We hypothesize that adapting BERTje to this domain through additional masked language modeling would enable the model to better capture the linguistic specificities of the archival texts.

This approach is intended to combine the strengths of pretraining and domain adaptation. However, as we discuss in the results section, the performance improvements from DAPT were limited, suggesting that continuing pretraining on a relatively small and noisy corpus—particularly one already used for fine-tuning—may not always yield substantial gains.

3 Results and Discussion

3.1 Quantitative Evaluation

We evaluate all models following the evaluation protocol used in Luthra et al. [11]. For each model, we report precision, recall, and F1 score at the micro level.

Table 1 summarizes the results. The BERTje + BiLSTM-CRF model reported in the original paper achieves a F1 score of 0.63, with precision in the range of approximately 0.68–0.70 and recall around 0.58–0.59. Our implementation of a Transformer model trained from scratch and combined with a CRF layer performs significantly worse, achieving a F1 score of 0.293, with a precision of 0.398 and a recall of 0.232. In contrast, our BERTje + CRF model achieves the highest overall performance, with a F1 score of 0.707, precision of 0.674, and recall of 0.743. The DAPT + BERTje + CRF model yields very similar results, with a F1 score of 0.705, precision of 0.674, and recall of 0.738, suggesting limited additional benefit from domain-adaptive pretraining in this setting.

Detailed evaluation results for each entity type under strict matching criteria are provided in Tables 3, 4, and 5.

Table 1: Precision, Recall, and F1 Scores for Different NER Models

Model	Precision	Recall	F1 Score
BERTje + BiLSTM-CRF (Paper)	∼0.68–0.70	∼0.58–0.59	0.63
Transformer + CRF (Ours)	0.398	0.232	0.293
BERTje + CRF (Ours)	0.674	0.743	0.707
DAPT + BERTje + CRF (Ours)	0.674	0.738	0.705

3.2 Analysis

Transformer Limitations The Transformer + CRF model, which is trained from scratch without any pretraining, performs significantly worse than the pretrained alternatives. This outcome is expected given the limited size of the annotated training corpus. Unlike models such as BERTje that begin fine-tuning with rich contextual embeddings learned from large-scale corpora, our Transformer must learn both linguistic representations and task-specific patterns from scratch using only a few thousand examples.

The model's architecture relies on self-attention to build contextualized token embeddings, but without sufficient data, it struggles to learn meaningful word associations, especially in a domain as linguistically complex and historically distant as 18th-century Dutch. As a result, the Transformer fails to generalize well and tends to overpredict the 'O' tag, leading to low recall and fragmented entity recognition. These results highlight the difficulty of training deep models from scratch in low-resource scenarios and confirm the importance of transfer learning for tasks involving historical or specialized domains.

BERTje + CRF Success The BERTje + CRF model achieves the best overall performance in our experiments, with a F1 score of 0.707, significantly outperforming both the baseline reported in the original paper and our untrained Transformer model. This strong result highlights the effectiveness of leveraging pretrained language models for named entity recognition, especially in low-resource and domain-specific contexts.

One possible reason for this improvement is architectural simplicity: unlike the model presented in the article, which adds a BiLSTM layer on top of BERTje, our model connects the CRF head directly to the contextual embeddings produced by BERTje. This may reduce overfitting and training instability, especially in a low-data regime. Additionally, minor differences in training procedure—such as the optimizer, batch size, learning rate, or the number of epochs—may contribute to improved generalization in our implementation. It is also possible that the random seed or the data preprocessing pipeline (e.g., tokenization, label alignment, or treatment of special tokens) differed slightly from that of the original authors, leading to small but impactful improvements.

Why DAPT Failed Despite the theoretical appeal of DAPT, our results show that it did not yield meaningful improvements over the standard BERTje + CRF model.

The dataset used for DAPT was the same corpus employed for fine-tuning, only stripped of annotations. While this aligns with typical DAPT methodology, it may have provided insufficient new linguistic information to significantly shift the model's internal representations. Moreover, the corpus is relatively small, historically noisy, and domain-narrow, limiting the effectiveness of masked language modeling as a signal for adaptation.

These results highlight that, while DAPT remains a promising tool in domain-specific NLP, its utility is highly sensitive to corpus size, domain distance, and training strategy.

4 Conclusion

This study evaluated several neural architectures for named entity recognition on Dutch colonial archival texts, focusing on the VOC testaments. We implemented and compared a Transformer + CRF model trained from scratch, a BERTje + CRF model leveraging pretrained Dutch language representations, and an additional variant with domain-adaptive pretraining.

Our findings confirm the critical role of transfer learning in low-resource and domain-specific settings. The Transformer model trained from scratch performed poorly due to limited data, while the BERTje + CRF model achieved strong results, even surpassing the baseline reported in prior work. These results reinforce the value of pretrained contextual embeddings for historical NER and point to the effectiveness of simple, well-regularized architectures.

Future work could further explore domain adaptation strategies and expand entity typologies to capture the full complexity of colonial records, while remaining attentive to the ethical implications of automated archival analysis.

References

[1] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gerrit Noord, and Malvina Nissim. Bertje: A dutch bert model. In *Proceedings of the 30th Benelux Conference on Artificial Intelligence and the 29th Belgian Dutch Conference on Machine Learning*, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186. Association for Computational Linguistics, 2019.

[3] Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named entity recognition and classification on historical documents: A survey. *arXiv preprint arXiv:2109.11406*, 2021.

[4] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics, 2020.

[5] Barry Hendriks, Paul Groth, and Marieke van Erp. Recognising and linking entities in old dutch text: A case study on voc notary records. In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, 2020.

[6] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. In *Proceedings of the 29th Conference on Neural Information Processing Systems*, pages 774–782, 2015.

[7] Charles Jeurgens and Michael Karabinos. Paradoxes of curating colonial memory. *Archival Science*, 20(3):199–220, 2020.

[8] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[9] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics, 2016.

[10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[11] Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. Unsilencing colonial archives via automated entity recognition. *arXiv preprint arXiv:2210.02194*, 2022.

[12] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[13] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[14] Joan M Schwartz and Terry Cook. Archives, records, and power: The making of modern memory. *Archival Science*, 2(1-2):1–19, 2002.

[15] Michel-Rolph Trouillot. *Silencing the Past: Power and the Production of History*. Beacon Press, 1995.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Table 2: IOB Tag Statistics per Document

IOB Tag(s)	Min	Mean	Max	n_unique_words
B-Organization	0.0	0.5	7	258
B-Organization,B-Place	0.0	0.0	1	2
B-Organization,I-Person	0.0	0.0	2	3
B-Organization,I-Place	0.0	0.0	2	14
B-Person	0.0	5.2	47	3501
B-Person,B-Place	0.0	0.0	1	2
B-Person,I-Place	0.0	0.0	1	9
B-Place	0.0	1.9	12	1081
I-Organization	0.0	1.4	25	777
I-Organization,B-Place	0.0	0.1	3	119
I-Organization,I-Person	0.0	0.0	7	9
I-Organization,I-Person,B-Place	0.0	0.0	1	1
I-Organization,I-Person,I-Place	0.0	0.0	1	1
I-Organization,I-Place	0.0	0.1	5	76
I-Person	0.0	10.0	80	8852
I-Person,B-Place	0.0	0.1	16	89
I-Person,I-Place	0.0	0.0	5	48
I-Place	0.0	2.5	104	1723
O	0.0	205.7	633	56915

Table 3: Detailed Evaluation Metrics for Transformer + CRF Model

Entity Label	Precision	Recall	F1-Score	Support
Organization	0.23	0.01	0.02	243
Organization,B-Place	0.00	0.00	0.00	31
Organization,I-Place	0.00	0.00	0.00	20
Person	0.40	0.28	0.33	2017
Person,B-Place	0.00	0.00	0.00	13
Person,I-Place	0.00	0.00	0.00	10
Place	0.41	0.19	0.26	594
Micro Avg	0.40	0.23	0.29	2928
Macro Avg	0.15	0.07	0.09	–
Weighted Avg	0.38	0.23	0.28	–

Table 4: Detailed Evaluation Metrics for BERTje + CRF Model

Entity Label	Precision	Recall	F1-Score	Support
Organization	0.38	0.39	0.38	243
Organization,B-Place	0.41	0.42	0.41	31
Organization,I-Place	0.45	0.25	0.32	20
Person	0.76	0.83	0.80	2017
Person,B-Place	0.35	0.62	0.44	13
Person,I-Place	0.00	0.00	0.00	10
Place	0.53	0.63	0.57	594
Micro Avg	0.67	0.74	0.71	2928
Macro Avg	0.41	0.45	0.42	–
Weighted Avg	0.67	0.74	0.71	–

Table 5: Detailed Evaluation Metrics for DAPT + BERTje + CRF Model

Entity Label	**Precision**	**Recall**	**F1-Score**	**Support**
Organization	0.33	0.45	0.38	243
Organization,B-Place	0.35	0.26	0.30	31
Organization,I-Place	0.33	0.05	0.09	20
Person	0.79	0.83	0.81	2017
Person,B-Place	0.60	0.46	0.52	13
Person,I-Place	0.00	0.00	0.00	10
Place	0.51	0.62	0.56	594
Micro Avg	0.67	0.74	0.70	2928
Macro Avg	0.42	0.38	0.38	–
Weighted Avg	0.68	0.74	0.71	–