https://doi.org/10.2352/issn.2168-3204.2024.21.1.2

Making sense of bureaucratic documents – Named entity recognition for state authority archives

Venla Poso, University of Jyväskylä; Jyväskylä, Finland; Mikko Lipsanen; The National Archives of Finland; Helsinki, Finland; Ida Toivanen; University of Jyväskylä; Jyväskylä, Finland; Tanja Välisalo; University of Jyväskylä; Jyväskylä, Finland.

Abstract

The usability and accessibility of digitised archival data can be improved using deep learning solutions. In this paper, the authors present their work in developing a named entity recognition (NER) model for digitised archival data, specifically state authority documents. The entities for the model were chosen based on surveying different user groups. In addition to common entities, two new entities were created to identify businesses (FIBC) and archival documents (JON). The NER model was trained by fine-tuning an existing Finnish BERT model. The training data also included modern digitally born texts to achieve good performance with various types of inputs. The finished model performs fairly well with OCR-processed data, achieving an overall F1 score of 0.868, and particularly well with the new entities (F1 scores of 0.89 and 0.97 for JON and FIBC, respectively).

Introduction

The National Archives of Finland started mass digitising state authority archives in 2019. The mass digitisation project will cover approximately 135 shelf kilometres of government documents [1]. Similar processes are in motion in archives all over the world, and the resulting challenges are shared globally. Accessibility and usability of the digitised data are the key elements for any digitisation process [2]. In order to achieve these goals and to make digitised archives useful for different user groups, the archival data needs to be enriched using different methods [3]. In this paper, we present our work with the application of named entity recognition (NER) for state authority archives.

Currently, the digitisation process at the National Archives of Finland begins with scanning after which these archival documents are post-processed, including extracting text with optical character recognition (OCR). This process improves access to the data but its usability for larger scale analysis still remains at a low level. The National Archives of Finland is currently in the midst of developing several deep learning models to be included in the digitisation process with the aim of improving the usability of the digital archives. In this paper, we propose a new NER model for state authority archives to improve their usability.

Named entity recognition (NER) is one of the most crucial steps in information retrieval from textual data [3]. In practice, NER models recognise words or sequences belonging to predetermined categories, such as 'person', 'date' or 'location'. Applications of NER in archival and historical domains have grown in number alongside digitisation efforts. For historians, the use of archives has been shown to rely on information retrieval tasks such as NER [4]. NER has benefits not only by itself but as a base for other machine learning tasks. Recognising named entities from unstructured text can enable information and document retrieval from a large collection of texts when, for example, the extracted information is used as a base for indexing or metadata creation [3]. In addition to indexing or metadata formation, NER can be used as a base for different data analysis methodologies or further data enrichment practices (e.g., entity linking, see [5] or relation extraction, see [6]).

In the development of NER for archival data lie some difficulties which are linked to the historical nature of the documents. In their survey, Ehrmann et al. [3] identify and generalise these challenges into four categories: domain heterogeneity, input noisiness, dynamics of language, and lack of resources. From these, we identified the first two as applying in our case as the state authority archives can include a variety of domains and the majority of archival data is noisy as a result of the OCR process.

In addition to presenting the NER model for Finnish archival data, our paper aims to share experiences and best practices for developing NER for the specific type of archival data that state authorities produce. As the project was executed in a multidisciplinary team in cooperation between two organisations, the National Archives of Finland and the University of Jyväskylä, we will also discuss the challenges of and best practices for such collaboration. Our team brought together people with backgrounds in history, and computer science.

Next, we will present the model development, starting with the choice of named entities and the annotation process, followed by a description of the model training, and finally, the model performance on the NER task.

The annotation process for state authority archives

The starting point of many NER applications has been the recognition of persons, organisations and locations, but the assortment of entity categories has expanded to fit different needs in different domains. For example, named entity processing has become widely popular in the biomedical domain, where categories can include entities such as proteins, genes or diseases [7][8]. Similarly, we considered the need for domain-specific entity categories in the archival domain in our model. User needs were mapped out beforehand with an online survey [9] for researchers using state authority archives. Based on this survey and a similar survey by the National Archives of Finland, aimed at different state authorities, the following named entities were chosen for implementation in the model development: person (PERSON), organisation (ORG), location (LOC), geopolitical location (GPE), product (PRODUCT), event (EVENT), date (DATE), nationality, religious and political groups (NORP), Finnish business identity code (FIBC), and journal number (JON). The first eight named entities were included in an existing model, TurkuNLP's NER model [10][11][12], which we used for pre-annotating the data (i.e.,

for creating pseudo-labels to the data [13]). The last two were completely new entities: FIBC is a permanent identifier which enables tracing businesses and organisations even when they change their name, and JON is a similarly permanent identifier for archival documents, which helps with tracking and correct dating of individual documents.

Our training data was selected from the readily available set of digitised state authority data sets. The data from the Ministry of Economic Affairs and Employment of Finland fitted our needs. In addition, we selected a variety of other archival data sources to compensate for some missing qualities from our primary dataset. These include articles from the Official Journal of Finland (Virallinen lehti), which is a journal for official state authority announcements, documents produced by an association of companies working in the environmental sector (Ympäristöyritysten liitto ry), as well as documents from the Chemical Industry Federation of Finland (Kemianteollisuus ry). In practice, this meant selecting numerous sources with variant cases of our two new entity groups, journal number and Finnish business identity code. The selection process was conducted by a professional archivist who had worked with the data before. As part of the data is sensitive and contains personal information, the annotation and model training had to be handled with special care. This also means that the training data will not be made publicly available.

Our objective was twofold: developing a NER model for state authority archives but also creating a method for handling sensitive archival data. Due to the sensitive data included in the archives, such as personal information, an environment tailored for sensitive data was used during the annotation process. This container (Apptainer in our case) based environment was designed so that it included coding packages necessary for preprocessing the data to make it more suitable for annotation purposes. Steps like changing the file type of the data from AltoXML to csv, filtering out redundant information from the files, and choosing files written primarily in Finnish were included in the preprocessing phase. Annotation was conducted with the IOB2 scheme (B tag for beginning of the word, I tag for following words which were inside the entity and O tag for words that were outside the annotation categories). Additionally, we included nested entities in the annotation process to improve future research perspectives and possibilities (see more [14]).

During our annotation phase, we discovered that the OCR-processed state authority data presented two main challenges. First, the noise generated by the OCR caused issues for annotation and model performance, which was on some level to be expected. A more complex and even unexpected obstacle presented itself in a form of in-domain language diversity. The formal state authority language, compared with data that has been previously used with Finnish NER, such as contemporary newspaper and magazine articles, blog posts, Wikipedia articles, and legal texts, was different in a way that presented unexpected questions about the actual meaning and boundaries of the categories. As state authorities can have complex organisational structures, which can go through several changes over time, it can be difficult to determine which words and phrases can be annotated as ORG. In addition, the PRODUCT entity was particularly challenging due to the broad variation in the types of 'things' it contained as well as the practical differences between commercial products and the outputs created by state authorities.

The two main challenges, OCR-noise, and state authority language, were met followingly:

To unify our practices in borderline cases, we created annotation guidelines that were refined by the annotators throughout the process. In practice, we had an online discussion platform, where both the annotators and model developers could discuss puzzling cases that they encountered and form unified instructions for prospective similar cases. The platform made it possible for all the participants to stay informed at all stages of the process. In addition, during the busiest annotation phase there were weekly meetings, which included all collaborators.

A multitude of possible solutions for OCR errors have been presented in previous studies. Solutions can be divided into two different approach groups: input correction (e.g., [15]) and adapting the model to OCR noise (e.g., [16]). Our approach belongs to the latter group. We aimed at providing examples of different possible OCR errors for the model by including words and phrases with minor OCR generated errors in the annotations (see Figure 1). We chose this approach due to the large quantity of data, which made manual correction of the training data impossible. Furthermore, the end goal of our model development is to create a NER model applicable for all National archives state authority data. These OCR errors included misrecognised, missing, or extra characters, as well as nearby words grouped together, or division of a word into several subwords [17].

Text (without noise)	Text (with noise)	Translation	Label
Dokumentin	okumentin	Document	O
1	1	1	B-JON
/	/	/	I-JON
234	234	234	I-JON
/	/	/	I-JON
2000	2000	2000	I-JON
mukaan	mukaan	presents (that)	O
Mikko	Mikko	Mikko	B-PERSON
Mallikas	M4llik4s	Mallikas	I-PERSON
on	onsdf	has	O
työskennellyt	työskennellyt	worked	O
yrityksen	yrityksen	(for the) company	O
1234567	1234567	1234567	B-FIBC
-	-	-	I-FIBC
8	8	8	I-FIBC
leivissä	leivissä	since	O
vuodesta	vuodesta	(the) year (of)	B-DATE
2000	2000	2000	I-DATE
.	.	.	O

Figure 1. *An example of the OCR noise included in the annotation.*

The state archival dataset was annotated by seven annotators, and to determine the level of agreement between all annotators we calculated the inter-annotator agreement score (Fleiss' kappa 0.84). If we follow the interpretation of Landis and Koch [18], our score falls on the highest level of agreement, 'almost perfect' (between 0.81 and 1.00).

The model development

We have built and published a new Finnish NER model in the HuggingFace platform. In addition to the data from the National Archives described previously, we also used other datasets for training a model that performs well with various types of input. The

data used for model implementation consisted of several datasets dating from different eras (listed in order of size):

1) Diverse state authority archival documents from Finnish public administration from 1970s - 2000s,
2) Turku OntoNotes Entities ("TurkuONE") Corpus from 2000s [12],
3) NewsEye dataset from 1850-1950 [19],
4) Finnish senate documents from 1916,
5) Finnish books from Project Gutenberg from the early 20th century and
6) theses from Finnish polytechnic universities from 2000s.

TurkuONE and NewsEye corpora were readily available with IOB2 annotations but were supplemented with missing named entity categories (i.e., FIBC, JON). For the datasets 4 - 6, all 10 named entities were annotated during the project. The state authority documents cover 69% and TurkuONE 23% of the whole data, while the remaining data sources (datasets 3–6) amount to 8% of the total. In total, the training data included over 128 000 annotated entities. The model was designed to work well for OCR-processed archival data, but training data also included modern, digitally born texts (dataset 2) in order to train a model that generalises well to various types of unseen input.

Table 1: Test results for the named entity recognition model for the 10 entity categories used in model training.

Entity group	Precision	Recall	F1 score
PERSON	0.90	0.91	0.90
ORG	0.84	0.87	0.86
LOC	0.84	0.86	0.85
GPE	0.91	0.91	0.91
PRODUCT	0.73	0.77	0.75
EVENT	0.69	0.73	0.71
DATE	0.90	0.92	0.91
JON	0.83	0.95	0.89
FIBC	0.95	0.99	0.97
NORP	0.91	0.95	0.93

The model was trained by fine-tuning an existing Finnish BERT model [20][21] for the named entity recognition task. The model was trained using a single NVIDIA RTX A6000 GPU. In the preprocessing stage, the input texts were split into chunks with a maximum length of 300 tokens. Tokenization was performed using the tokenizer for the cased FinBERT base model. The model was trained for 10 epochs (with patience of 2 epochs) with the batch size of 24. AdamW was used as optimizer (with betas (0.9,0.999) and epsilon 1e-08) along with linear scheduler and learning rate of 2e-05. For the scheduler, the number of warmup steps was calculated as the amount of training data divided by five, and the number of training steps as the amount of training data times epochs. Dropout of 0.3 was used to regularise the model. In addition to the parameters of the model classification layers, the parameters of the base model were also tuned during training in order to prepare the model to better recognise both OCR-processed text and the type of language used in the state authority documents. When tested with a test dataset containing documents from the same domains as the training data, the model achieved a mean (non-weighted) F1 score of 0.868. There was significant variation in the results for different entity classes, with the best (FIBC) F1 score of 0.97 and the worst (EVENT) 0.71. Table 1 presents the results for each entity group.

Discussion and conclusions

In this paper, we have presented a NER model for improving the usability of state authority archives. The test results show that the F1 score, which is based on a harmonic mean of precision and recall, the model performed fairly well, reaching a score of 0.868. For two entities, EVENT and PRODUCT, the model performance was significantly lower than for the others. The reasons for low scores can be explained in a few ways.

The occurrences of the EVENT entity in the data were quite rare, which probably explains the lower level of performance. The PRODUCT entity, which proved challenging in the annotation phase due to the differences between the training data of the original model and the state authority archives used here, achieved lower results (F1 0.75) here as well. This suggests a need for reconsidering the definition and usefulness of the PRODUCT entity, in particular with historical data. The two new entities, JON (F1 0.89) and FIBC (F1 0.97) performed well, most probably because both of them have very consistent formatting. This is encouraging for their future usefulness for different archive user groups.

The regular discussions between the two organisations involved were essential for successful collaboration. In particular, the online discussion platform proved important in providing a channel for discussion and support between meetings enabling the work to continue proficiently. Data security policies and differences in communication infrastructures can sometimes hinder collaboration across organisational borders. However, we found it critical to cross these potential obstacles to enable the flow of information between collaborators.

Despite the success of our NER model, it still has many limitations that should be taken into account when considering its broader applicability. While more training data was used than in previous Finnish NER models, a bigger training dataset would most probably help to improve the results further.

In addition to the amount of data, its diversity is also a key factor in improving the generalisability of the model. We used state authority data only from the Ministry of Economic Affairs and Employment of Finland. Including documents produced by other state authorities in the training data would, however, contribute to the diversity of the language and named entities contained in the dataset. Regarding the OCR-noise, the text content of our state authority training data has been recognised using a specific software, Tesseract OCR, which likely biases the model to better handle the type of noise this OCR engine is likely to produce.

Due to the confidential nature of the state authority data, our full training dataset cannot be made publicly available, which sets a constraint on the reproducibility of the results. We will, however, publish the guidelines that were used for data annotation.

To better assess and solve some of the limitations listed above, further testing and development of the model is needed. Future work should include comparisons between this model and other existing NER models for Finnish language data with both archival data and digitally born texts. Further ahead, the model needs to be

implemented with a graphical user interface in connection to other tools provided by the National Archives of Finland, where it could be used for automatic metadata creation and to improve existing search tools.

References

[1] T. Hölttä and V. Kajanne(2020). No more new archive buildings – mass digitisation and retroactive digitisation improve the accessibility of material. In Nuorteva, J. & Happonen, P. (eds.), The National Archives of Finland Strategy 2025. https://kansallisarkisto.fi/documents/141232930/153230445/KA_Strategy_2025_eng.pdf/.

[2] Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. "Archives and AI: An Overview of Current Debates and Future Perspectives." J. Comput. Cult. Herit. vol. 15, no. 1, 2021. https://doi.org/10.1145/3479010

[3] M. Ehrmann, A. Hamdi, E. Linhares Pontes, M. Romanello and A. Doucet, "Named Entity Recognition and Classification in Historical Documents: A Survey," ACM Comput. Surv. vol. 56, iss. 2, Article 27, Sept 2023. Available: https://doi.org/10.1145/3604931.

[4] W. M. Duff and C. A. Johnson, "Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives", The Library Quarterly, vol. 72, iss. 4, pp. 472–496, Oct 2002. Available: https://www.jstor.org/stable/40039793

[5] S. Tedeschi, S. Conia, F. Cecconi, R. Navigli, "Named Entity Recognition for Entity Linking: What Works and What's Next", Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 2021, pp. 2584–2596. Available: https://aclanthology.org/2021.findings-emnlp.220.pdf

[6] Z. Nasar, S. W. Jaffry and M. K. Malik, "Named Entity Recognition and Relation Extraction: State-of-the-Art", Comput Surveys, vol. 54, iss. 1, Article 20, Feb 2021. Available: https://doi.org/10.1145/3445965

[7] R. R. Villarreal Goulart, V. L. Strube de Lima and C. Castellã Xavier, "A Systematic Review of Named Entity Recognition in Biomedical Texts", J Braz Comput Soc, vol. 17, iss. 2, pp. 103–116. June 2011. Available: https://doi.org/10.1007/s13173-011-0031-9

[8] H. Cho and H. Lee, "Biomedical named entity recognition using deep neutral networks with contextual information", BMC Bioinformatics, vol. 20, Article 735, 2019. Available: https://doi.org/10.1186/s12859-019-3321-4

[9] V. Poso, T. Välisalo, I. Toivanen, A. Holmila and J. Ojala, "Untapped data resources. Applying NER for historical archival records of state authorities", DHNB Publications, vol. 5, no. 1, pp. 55–69, Oct 2023. Available: https://doi.org/10.5617/dhnbpub.10650

[10] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, "A Broad-coverage Corpus for Finnish Named Entity Recognition", Proceedings of The 12th Language Resources and Evaluation Conference (LREC'2020), Marseille, France, pp. 4615–4624. Available: https://aclanthology.org/2020.lrec-1.567/

[11] J. Luoma and S. Pyysalo, "Exploring cross-sentence contexts for named entity recognition with BERT", arXiv preprint, June 2020. Available: https://doi.org/10.48550/arXiv.2006.01563

[12] J. Luoma, L. Chang, F. Ginter, S. Pyysalo, "Fine-grained Named Entity Annotation for Finnish", Proceedings of the 23rd Nordic Conference on Computational Linguistics, NoDaLiDa, Reykjavik, Iceland, 2021, pp. 135–144. Available: https://aclanthology.org/2021.nodalida-main.14

[13] D. H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." In Workshop on challenges in representation learning, ICML, vol. 3, no. 2, pp. 896, 2013. https://www.kaggle.com/blobs/download/forum-message-attachment-files/746/pseudo_label_final.pdf.

[14] J. R. Finkel and C. D. Manning, "Nested Named Entity Recognition", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 2009, pp. 141–150. Available: https://aclanthology.org/D09-1015

[15] V.-N. Huynh, A. Hamdi and A. Doucet, "When to Use OCR Post-Correction for Named Entity Recognition?", Digital Libraries at Times of Massive Societal Transition, 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, 2020, pp. 33–42. Available: https://doi.org/10.1007/978-3-030-64452-9_3

[16] E. Boros, A. Hamdi, E. Linhares Pontes, L. A. Cabrera-Diego, J. G. Moreno, N. Sidere and A. Doucet, "Alleviating Digitization Errors in Named Entity Recognition for Historical Documents", Proceedings of the 24th CoNLL, 2020, pp. 431–441. Available: https://doi.org/10.18653/v1/2020.conll-1.35

[17] E. Soper, S. Fujimoto and Y.-Y. Yu, "BART for post-correction of OCR newspaper text", Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), 2021, pp. 284–290. Available: https://aclanthology.org/2021.wnut-1.31.pdf

[18] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data", Biometrics, vol. 33, iss. 1, pp. 159–174, Mar 1977. Available: https://doi.org/10.2307/2529310

[19] A. Hamdi, E. Linhares Pontes, E. Boros, T. T. H. Nguyen, G. Hackl, J. G. Moreno and A. Doucet, "A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers", Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2328–2334, July 2021. Available: https://doi.org/10.1145/3404835.3463255

[20] Available: https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1

[21] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter and S. Pyysalo, "Multilingual is not enough: BERT for Finnish", arXiv preprint, Dec 2019. Available: https://doi.org/10.48550/arXiv.1912.07076

Author Biographies

Venla Poso (MA, MSc), is a doctoral researcher at the University of Jyväskylä. At the moment Poso works in a research infrastructure which integrates DARIAH-FI and FIN-CLARIN networks. Her research interests include developing and applying AI models in digital humanities research, particularly with textual data.

Mikko Lipsanen has a MSc in Data Science from the University of Helsinki, Finland. His work in the National Archives of Finland focuses on the use of Machine Learning to improve the processing and accessibility of digitized data.

Ida Toivanen (MSc) is a doctoral researcher at the University of Jyväskylä, working on deep learning based solutions that can be applied into digital humanities in the infrastructure project FIN-CLARIAH. Toivanen is interested in learning all about DL – from NLP and computer vision to multimodal systems.

Tanja Välisalo (PhD) is a researcher at the University of Jyväskylä, where she also teaches digital research methods. Välisalo is involved in developing digital research tools for various types of data from digital and digitized documents to online discussions and online chat data. Välisalo is affiliated with the Centre of Excellence in Game Culture Studies funded by the Research Council of Finland.