Named Entity Recognition for digitised archival documents in German

Nele Garay[1], Mahsa Vafaie[1,2] and Harald Sack[1,2]

[1]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2]Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology (KIT), Germany

Abstract

This paper presents an experiment that evaluates the effectiveness of two different Named Entity Recognition (NER) tools at extracting entities directly from the output of an Optical Character Recognition (OCR) workflow. The authors initially developed a test dataset comprising both raw and corrected OCR outputs, which were manually annotated with tags for Person, Location, and Organisation. Subsequently, they applied each NER tool to both the raw and corrected OCR outputs, evaluating their performance by comparing the precision, recall, and F1 scores against the manually annotated data.

Keywords

Named Entity Recognition (NER), Optical Character Recognition (OCR), Digital Cultural Heritage

1. Introduction

In the field of Natural Language Processing (NLP), Named Entity Recognition (NER) is an essential task that facilitates extraction of structured data from unstructured text [1]. This function is especially beneficial when processing text acquired through Optical Character Recognition (OCR), which transforms digitised documents into editable and searchable formats [2]. Nevertheless, OCR-generated text frequently includes noise and errors due to recognition inaccuracies, presenting distinct challenges for NER systems.

One of the applications in which OCR and NER technologies commonly meet is in the context of digitalisation, where large amounts of text are produced through the OCR process, and information is to be extracted from the body of unstructured text. *"Themenportal zur Wiedergutmachung nationalsozialistischen Unrechts"* [1] [*Thematic Portal for Compensation of National Socialist Injustice*] is a digitalisation project that aims to create an information system for contextualisation of historical knowledge derived from collections of documents, records, and materials directly linked to the compensation process for the atrocities of the National Socialist regime in Germany. The first step in this digitalisation initiative is converting document images collected from archives across Germany, into machine-readable formats, through Text Recognition technologies [3]. The workflow then involves extracting information from the OCR text, designing and populating ontologies [4, 5], and linking the extracted entities with external sources to build the "Wiedergutmachung Knowledge Graph". As part of the information extraction pipeline for this project, Named Entity Recognition is considered a powerful tool, for identification of entities within vast amounts of unstructured or semi-structured data (e.g. tables and forms), either as a stand-alone tool or in combination with other methods such as regular expressions or end-to-end information extraction with LLMs, to increase reliability and confidence in the output.

Despite advances in both OCR and NER technologies, the combination of these two techniques remains a complex and evolving area of research. By assessing the impacts of OCR noise on NER and comparing the performance of two different NER tools on OCR text, this study aims to contribute

EKAW-24: 24th International Conference on Knowledge Engineering and Knowledge Management, November 26-28, 2024, Amsterdam, Netherlands
nele.garay@fiz-karlsruhe.de (N. Garay); mahsa.vafaie@fiz-karlsruhe.de (M. Vafaie); harald.sack@fiz-karlsruhe.de (H. Sack)
0009-0007-9697-3630 (N. Garay); 0000-0002-7706-8340 (M. Vafaie); 0000-0001-7069-9804 (H. Sack)

[1]https://is.gd/bundesfinanzministeriumwgm

to the goal of making vast amounts of noisy textual data more accessible. This assessment is only possible through the creation of gold standard datasets. In this work, we present two openly available datasets from the collection of *Wiedergutmachung* documents, manually annotated with the most commonly used NER tags. Moreover, we compare two open-source NER models, a general model and a domain-specific one, to provide insights and discussions on the efficacy of these tools for NER with noisy OCR textin a prticular domain. In Section 2, a brief overview of similar attempts to assess NER quality on OCR generated text is provided. Section 3 introduces the two datasets and the NER tools with which the experiments have been conducted. Section 4 presents the results and a comparison between the two models and across different tags, with noisy and clean transcripts. Section 5 concludes the paper by emphasising the discussion points and proposing future directions for the combination of OCR and NER techniques.

2. Related Work

Named Entity Recognition (NER) tries to identify words and expressions that belong to particular categories of Named Entities. Most commonly the categories include names of persons, location and organisation. Identifying Named Entities can help with finding specific documents in collections[6].

Optical Character Recognition (OCR) systems transcribe pictures and scans of text documents. Due to different factors such as bad quality of digitisation, or special fonts that cannot be easily recognised by OCR engines, OCR- generated texts often contain errors, that can potentially have a negative impact on the output of NER. This challenge has been studied and addressed by many researchers [7]. In [8] the authors tested different tools of extracting person names from historical OCRed documents. When comparing with hand annotated texts they found that OCR mistakes in word order had a bigger impact in NER results than character recognition errors. Rodriquez et al. [2] compared NER tools on OCRed text of historical Holocaust related documents from the European Holocaust Reasearch Infrastructure (EHRI documents), which include among others newspaper articles, victim testimonies and diplomatic reports. They noted that the correction of the OCR text does not increase the performance of their NER tools by a significant amount. Ruokolainen et al. [9] trained two NER-models on OCRed Finnish language newspaper text. Evaluation results show F1-scores above 0.72 for Location and Person tags. To increase the score for Organisation tags, a nested entity approach was used which resulted in an F1-score of 0.44. Koudoro-Parfait et al. [10] tested the impact of different OCR systems on NER evaluation of french novels. They found a negative correlation between OCR quality and NER quality, but missing blank spaces, faulty first characters, and wrong word order are the OCR errors of the greatest impact. Hamdi et al. [7] studied five types of OCR errors and their impact on the performance of NER. They found that segmentation errors (wrong word order) and errors in the first character had a strong impact on the performance of NER.

Recent advances in NER have led to the development of systems that are pre-trained on large amounts of contemporary datasets and are ready for use in various languages [11, 12]. These tools leverage state-of-the-art techniques, including Transformer-based models and deep learning architectures, to enhance their performance across different linguistic contexts [13]. Two examples for such NER tools are described in more details below.

Flair [14] proposes a solution to challenges and problems with contextualised embeddings of words by using a simple, unified interface for word embeddings. Flair also provides pre-trained models for different languages and use-cases, including the NER-tagger for German language.

European Holocaust Research Infrastructure (EHRI) has recently developed a single multilingual NER model from a multilingual dataset of Holocaust related documents. With this dataset, the multilingual Transformer-based masked language model XML-RoBERTa-large has been fine-tuned. The EHRI-NER model performs well on Holocaust specific datasets with an F1-score above 0.80 [15].

In this study we will use these two pre-trained NER systems mentioned above for our experiments on historical data from Germany between 1950s and 1980s.

3. Datasets and Experiments

The dataset used in this work consists of text files acquired with Optical Character Recognition (OCR) from one of the document collections of the *Wiedergutmachung* project. This collection, called *"Bundeszentalkartei"* [*Central Federal Index for Compensation*] or *BZK* in short, is a central card file of most of applications for compensation in the Federal Republic of Germany. These card files are in the form of pre-printed index cards, filled in either by typewriter or by hand.

The dataset collected for the evaluation consists of the OCR transcripts from 135 documents from the BZK collection, here referred to as BZK, and the manually corrected version of the same OCR transcripts, here referred to as BZK-GT. Both datasetes are collected such that they do not fall under strict data privacy restrictions and are therefore, openly available [2]. A Transformer-based OCR model from Transkribus [3], called TextTitan, has been used for text recognition, since the documents contain a mix of machine-printed and handwritten text, and Transformer-based OCR models have shown to perform well on documents with multiple text types [16, 17].

Both BZK and BZK-GT datasets have been manually annotated with entity labels by one annotator. Before the manual annotation, the documents were tokenised as follows: First, all the dots and space characters that occur more than three times were discarded and replaced "/" with a space character. We also replaced abbreviations (e.g. *geb.* for *geboren*, *str.* for *straße*, *verst.* for *verstorben*) with their full form for better readability and recognition by the NER model. After these pre-processing steps, spaCy core[4] was used for tokenisation.

The named entity models we used contain the entity classes Person (PER), Location (LOC) and Organisation (ORG). For the manually created NER ground truth for both datasets, the NER classes are defined as follows:

- **PER** includes persons' first and last names without titles.
- **LOC** includes country, state, city and street names, as well as names of deportation and concentration camps.
- **ORG** includes names of governmental offices (*Entschädigungsämter* [*Offices for Compensation of National Socialist Injustice*]).

All datasets are created in the CoNLL 2003 format [18].

Since many organisation names in German also contain city names, we used a nested entity approach [19] when labelling organisations, i.e., a token can be both part of an organisation entity and a location entity. The phrase *"Entschädigungsamt Berlin"* is therefore tagged as:

```
Entschädigungsamt   O   B-ORG
Berlin   B-LOC   I-ORG
```

Tokens in the BZK dataset with OCR errors only get an NER tag if a maximum of two characters deviate from the original word (e.g., *Lonkom* instead of *London*) or if the entity is within a string next to some wrongly identified characters (e.g., *Londonpsc* instead of *London*).

Because the structure of the text on the cards is not always line by line, the OCR text sometimes has a wrong order. This led to challenges during the annotation process, especially for the BIO tagging, since without information about the document layout it was unclear if two entity tokens that appeared right next to each other belonged to the same entity or not. Another problem during annotating existed for multiple word entities that had been separated by the wrong word order. As a solution, while manually tagging the BZK dataset, the OCR text was used as a stand-alone text and the basis for the BIO tagging by itself, without information about the image and document layout. This led to a different BIO-tagging and fewer ORG-tags (23.65% less) in the BZK dataset, compared to the BZK-GT dataset, which better adheres to the document structure and layout.

[2] https://github.com/ISE-FIZKarlsruhe/Wiedergutmachung/tree/main/NER
[3] https://www.transkribus.org/de
[4] https://huggingface.co/spacy/en_core_web_sm

For the NER task two models were used: the German language model from Flair, called `ner-german-large`[5] as a general model, and the EHRI-NER model[6] which is trained on multilingual (including Czech, German, English, French, Hungarian, Dutch, Polish, Slovak, Yiddish) Holocaust related textual data, as a domain-specific tool. The results of these experiments and a discussion of the results follow in the next section.

4. Results and Discussion

The results of the NER evaluation using the two different NER tools are summarised in Table 1 for both the BZK and BZK-GT datasets. The evaluation indicates that the BZK-GT dataset achieves higher F1-scores compared to the BZK dataset. Both NER tools exhibit poorer performance on the noisy text generated by the OCR system, highlighting the need for OCR post-processing in the NER pipeline for raw OCR text.

NER tool	Dataset	Class	Precision	Recall	F1	Found Tags (TP+FP)	Tags in the Gold Standard
Flair-ner-de	BZK-GT	PER	0.97	0.84	0.90	416	475
		LOC	0.97	0.76	0.85	911	1152
		ORG	0.63	0.35	0.45	229	406
		Total	0.92	0.70	**0.79**	1556	2033
	BZK	PER	0.83	0.81	0.82	440	447
		LOC	0.96	0.73	0.83	835	1084
		ORG	0.51	0.31	0.39	193	310
		Total	0.86	0.68	**0.76**	1468	1841
EHRI-NER	BZK-GT	PER	0.86	0.96	0.90	530	475
		LOC	0.89	0.78	0.83	1011	1152
		ORG	0.83	0.44	0.58	217	406
		Total	0.87	0.75	**0.81**	1758	2033
	BZK	PER	0.70	0.93	0.80	589	447
		LOC	0.83	0.77	0.80	996	1084
		ORG	0.75	0.32	0.45	134	310
		Total	0.78	0.73	**0.76**	1719	1841

Table 1: Results of NER evaluation using the two different NER tools, on raw OCR text (BZK) and corrected OCR (BZK-GT)

Another observation from the table is that, across both datasets and models, the F1-scores for ORG entities are significantly lower than those for other tags. However, while the evaluations of Flair and EHRI-NER yield similar F1-scores for PER and LOC tags, the EHRI model achieves higher F1-scores for ORG entities compared to Flair. This results in higher overall scores for the EHRI-NER model. This

[5] https://huggingface.co/flair/ner-german-large
[6] https://huggingface.co/ehri-ner/xlm-roberta-large-ehri-ner-all

discrepancy can be attributed to the fact that the most prominent organisation types mentioned on BZK cards are now considered historical and most no longer exist. Consequently, language models trained on non-historical data struggle to recognise these historical organisations. Therefore, EHRI-NER, which is fine-tuned on historical data from the same period, outperforms the general pre-trained model in recognition of ORG entities in this dataset.

5. Conclusion and Future Work

In this study, we compared the performance of two Named Entity Recognition (NER) tools, the Flair German model and the EHRI-NER model, on German historical OCRed text, to determine if the Holocaust-specific EHRI-NER model outperforms the general model on our dataset, and to assess the impact of OCR noise on NER quality, using the two aforementioned models. A significant contribution of this work is the creation of two datasets, BZK and BZK-GT, which involved the manual annotation of raw and corrected OCR texts, as well as the evaluation of NER predictions from both models.

Our findings confirm that OCR errors degrade the quality of NER predictions for both models. However, the EHRI-NER model demonstrated strong performance on our datasets, particularly in recognising historical Organisations, in comparison to the Flari NER German model.

One of the primary challenges encountered during our experiments was the annotation of raw OCR text with entity labels. To address this challenge, developing comprehensive guidelines for annotation could significantly streamline this process. Such guidelines would not only speed up the annotation phase, but also enhance the consistency and comparability of annotated datasets, and provide the possibility to engage multiple annotators in the process, thereby improving the overall quality of future research.

In future research, fine-tuning the models using our specific datasets could potentially enhance the NER results we have achieved. Additionally, conducting experiments with other Large Language Models would provide valuable comparative insights. This approach could help identify the most effective models for our tasks and further improve the robustness and accuracy of NER predictions.

Acknowledgments

This work is funded by the German Federal Ministry of Finance (*Bundesministerium der Finanzen*).

References

[1] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (2007) 3–26.

[2] K. J. Rodriquez, M. Bryant, T. Blanke, M. Luszczynska, Comparison of named entity recognition tools for raw ocr text., in: Konvens, 2012, pp. 410–414.

[3] M. Vafaie, J. Waitelonis, H. Sack, Improvements in handwritten and printed text separation in historical archival documents, in: Archiving Conference, volume 20, Society for Imaging Science and Technology, 2023, pp. 36–41.

[4] M. Vafaie, O. Bruns, N. Pilz, D. Dessí, H. Sack, Modelling archival hierarchies in practice: Key aspects and lessons learned (2021).

[5] M. Vafaie, O. Bruns, N. Pilz, J. Waitelonis, H. Sack, Courtdocs ontology: Towards a data model for representation of historical court proceedings, in: Proceedings of the 12th Knowledge Capture Conference, 2023, pp. 175–179.

[6] A. Mikheev, M. Moens, C. Grover, Named entity recognition without gazetteers, in: Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999, pp. 1–8.

[7] A. Hamdi, E. L. Pontes, N. Sidere, M. Coustaty, A. Doucet, In-depth analysis of the impact of ocr errors on named entity recognition and linking, Natural Language Engineering 29 (2023) 425–448.

[8] T. L. Packer, J. F. Lutes, A. P. Stewart, D. W. Embley, E. K. Ringger, K. D. Seppi, L. S. Jensen, Extracting person names from diverse and noisy ocr text, in: Proceedings of the fourth workshop on Analytics for noisy unstructured text data, 2010, pp. 19–26.
[9] T. Ruokolainen, K. Kettunen, Name the name-named entity recognition in ocred 19th and early 20th century finnish newspaper and journal collection data., in: DHN, 2020, pp. 137–156.
[10] C. Koudoro-Parfait, G. Lejeune, G. Roe, Spatial named entity recognition in literary texts: What is the influence of ocr noise?, in: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, GeoHumanities '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 13–21. URL: https://doi.org/10.1145/3486187.3490206. doi:10.1145/3486187.3490206.
[11] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, arXiv preprint arXiv:1910.11470 (2019).
[12] K. Pakhale, Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges, arXiv preprint arXiv:2309.14084 (2023).
[13] M. Monteiro, C. Zanchettin, Optimization strategies for bert-based named entity recognition, in: Brazilian Conference on Intelligent Systems, Springer, 2023, pp. 80–94.
[14] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use framework for state-of-the-art nlp, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.
[15] M. Dermentzi, H. Scheithauer, Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools, in: Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024, 2024, pp. 18–28.
[16] P. B. Ströbel, T. Hodel, W. Boente, M. Volk, The Adaptability of a Transformer-Based OCR Model for Historical Documents, in: Intl. Conf. on Document Analysis and Recognition, Springer, 2023, pp. 34–48.
[17] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, Trocr: Transformer-based optical character recognition with pre-trained models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13094–13102.
[18] E. F. Sang, F. De Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, arXiv preprint cs/0306050 (2003).
[19] J. R. Finkel, C. D. Manning, Nested named entity recognition, in: Proceedings of the 2009 conference on empirical methods in natural language processing, 2009, pp. 141–150.