What's in an entity? Exploring Nested Named Entity Recognition in the Historical Land Register of Basel (1400-1700).

Long Paper proposal to the DH Benelux 2024

Prada Ziegler, Ismail
University of Bern, Switzerland; University of Basel, Switzerland
https://orcid.org/0000-0003-4229-8688

Introduction

In the late medieval and early modern period, there was a notable surge in the documentation of property ownership, sales, and rents. In the context of the city of Basel, these records were compiled in the early 20th century into the Historical Land Register[1], comprising source excerpts closely aligned with the original texts. This repository of information not only provides valuable insights into economic trends but also serves as a significant resource for understanding the social structure of the time. The pre-1800 collection encompasses approximately 120,000 entries of transactions (e.g. sales, inheritance, confiscations). We particularly focus on 60,000 documents dating from 1400 to 1700.[2] The magnitude of this archival compilation poses challenges for research inquiries and suggests that there is value in a machine-learning approach to extracting relevant information from the documents.

A first step for other information extraction methods, such as extracting events, is often the annotation of entities (Li et al., 2022). In previous experiments, I investigated the possibility of detecting named entities in the historical tower books of Bern (Hodel et al., 2023). An additional challenge arises from the documents within the historical land records: about 60% of all entity mentions therein are nested within other mentions (see Table A for an example). Nested Named Entity Recognition (NNER) is a well-known challenge in Natural Language Processing (Wang et al., 2022), although it receives much less attention than so-called "flat" Named Entity Recognition. For our purpose, standard NER recognition is not sufficient, as we would miss relations and events which are embedded within entity mentions. In this proposal, I report the findings of my experiments applying an established NNER architecture to the Historical Land Records of Basel compared to a method I developed.

Methodology

The documents were automatically transcribed using Handwritten Text Recognition with an average error rate of 3.6%. We annotated 571 excerpts with the BeNASch-system (Prada Ziegler et al., 2024), an annotation framework developed for pre-modern German texts,

---

1 https://www.staatsarchiv.bs.ch/benutzung/recherche/suche-gedruckte-kataloge/historisches-grundbuch.html (Visited 29.01.2024)

2 This experiment was conducted as part of the historical research project "Economies of Space" (https://dg.philhist.unibas.ch/de/bereiche/mittelalter/forschung/oekonomien-des-raums/).

which is inspired by the ACE2005 guidelines. ACE[3] is a common benchmark for NNER in modern English (Wang et al., 2022).
While our annotations are more nuanced, this paper focuses on detecting entity boundaries as well as classifying entity types. Targets are entity mentions of the types “person” (PER), “organization” (ORG), “location” (LOC) and the “heads” of those entities (e.g. the proper name inside the mention of a person).
To get a vectorized representation of the strings, I trained a character-based language model with the Python NLP framework Flair (Akbib et al., 2018; Akbib et al., 2019), fine-tuning a general modern German model provided by the Flair framework, using slightly preprocessed texts from the Historical Land Records.[4] Flairs character embeddings have previously proven to perform especially well with pre-modern German language (Hodel et al., 2023).
The initial approach I tested for the recognition task uses a classic BiLSTM architecture with a modified CRF layer to detect nested annotations using a “second-best” strategy developed by Shibuya and Hovy (2020). I chose this architecture because an implementation using Flair language models was publicly available and it scored state-of-the-art results compared to similar systems (Wang et al., 2022).
The second “architecture”, developed by myself, is a simple recursion strategy using a typical Flair model trained to annotate flat NER tags. This model is trained on modified training data, where each annotated span in a document is represented as a sample in the training data (see Table A).

Sample Text	Sample Annotation
Item Jacob Böglein, der Bader verkauft [...]	[“Jacob Böglein, der Bader” / PER]
Jacob Böglein, der Bader	[“Jacob Böglein” / Head], [“der Bader” /PER]
der Bader	[“Bader” / Head]

Table A: The span “Item Jacob Böglein, der Bader verkauft…” (engl. “Likewise, Jacob Böglein the barber sells…”) is expanded into three samples in the training data. Because heads may not contain entities, they do not spawn further samples.

On inference, the system starts annotation on the document level. Whenever an entity mention is found, the span of that mention is annotated until no more annotations are found or only Head annotations remain (which can not contain other mentions). We understand this as a “recursive” annotation system.

Results

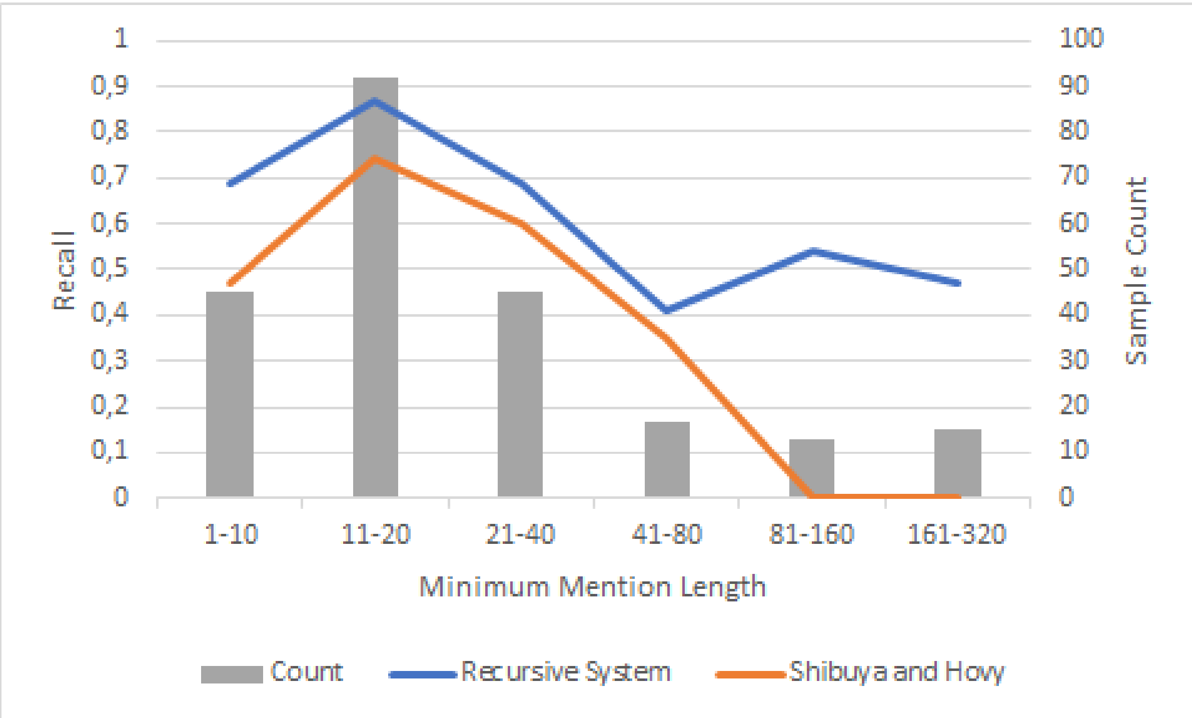
	Shibuya and Hovy 2020		**Recursive System**	
Span Type	**Precision**	**Recall**	**Precision**	**Recall**
PER	0.76	0.81	0.85	0.85
ORG	0.75	0.71	0.78	0.82

[3] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf (Visited 29.01.2024)
[4] The model as well as all accompanying scripts will be released on huggingface.co and Github respectively at a later date.

LOC	0.64	0.52	0.73	0.71
Head	0.8	0.84	0.84	0.85
Weighted Average	0.76	0.77	0.82	0.82

Table B: Precision and recall scores of both systems in comparison. Scores represent strict matching, meaning boundaries as well as categorization of the predicted annotation must match to be considered a true positive. Partial matches are counted as false positives.

The results of the experiments are shown in Table B. The recursive system performs better by each metric in each category, sometimes by a wide margin. Investigating the largest margin, the recall on location tags may help our understanding of the results. In Graph C, the location recall is compared between both systems. While the recursive system performs better in general, the largest difference can be observed in recognizing particularly long annotation spans. The system by Shibuya and Hovy struggles to recognize any annotations longer than 80 characters. These are mostly annotations on document level. An evaluation by depth shows that the recursive system recognizes 64.1% of all location mentions on document level, while the other system only recognizes 30.77%. These annotations are especially important for further historical analysis as those are usually the full descriptions of houses in the documents.
It is difficult to put the resulting numbers into context due to the lack of comparable experiments. Compared to the flat annotation experiments with the tower books of Bern based on similar material (Hodel et al., 2023) these numbers are exceptionally well, especially considering the increased difficulty of the task itself.



Graph C: Comparison of the recall scored by both systems according to maximum length of the annotation on location annotations. Reading example: Location annotations longer than 80 characters but shorter or equal to 160 characters in length could not be found by Shibuya and Hovy. At the same time, the recursive System recognized 54% of those annotations.

Conclusion

Nested entities in pre-modern German-language land records can be detected very reliably. While it comes as a surprise that a state-of-the-art architecture performs worse than a simple recursive application using a traditional NER model with some modified training material, this can probably be attributed to the special setting of the experiment: For one, appr. 30’000 tokens training material is an extremely small corpus to train on, and second, the architecture by Shibuya and Hovy is usually used with richer embeddings than just Flair, such as BERT embeddings and stacking them.
Finally, I would attribute the high scores achieved in this experiment to a large part to the domain of the utilized texts. The documents collected in the land records often follow a strict structure. Such repeating patterns are easily picked up by machine learning systems. Furthermore, the frequent repetition of names due the identical geographical origin of the registers is supportive. How well a model like this can generalize to other domains or land registers from other cities should be subject to further research.
These findings lay the groundwork to perform further information extraction tasks on historical mass data. Reliable detection of named entities will be important to the tasks of relationship and event extraction in particular.

Bibliography

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, June). 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)* (pp. 54-59).

Akbik, A., Blythe, D., & Vollgraf, R. (2018, August). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638-1649).

Hodel, T., Prada Ziegler, I., & Schneider, C. (2023). Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern. Presented at the Digital Humanities 2023. Collaboration as Opportunity (DH2023), Graz, Austria, Jun. 2023.

Li, Q., Li, J., Sheng, J., Cui, S., Wu, J., Hei, Y., Peng, H., Guo, S., Wang, L. & Philip, S. Y. (2022). A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21.

Prada Ziegler, I., Weber D. & Schneider, C. (2024). BeNASch - a Common Semantic Annotation Framework for Early Modern German. Presented at the Open Up Digital Editions 2024 (OpenUpDSE24), Zurich, Switzerland, Jan. 2024.

Shibuya, T., & Hovy, E. (2020). Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, *8*, 605-620.

Wang, Y., Tong, H., Zhu, Z., & Li, Y. (2022). Nested named entity recognition: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(6), 1-29.