Information Extraction from German Medieval Charters Abstracts

Sandy Aoun[1†], **Varvara Arzt**[2,3], **Daniel Luger**[1], **Georg Vogeler**[1]

[1]Department of Digital Humanities, University of Graz, Austria
{sandy.aoun, daniel.luger, georg.vogeler}@uni-graz.at

[2]Faculty of Informatics, Vienna University of Technology, Austria
varvara.arzt@tuwien.ac.at

[3]Digital Age Research Center, University of Klagenfurt, Austria

The large-scale transformation of historical records into machine-readable formats has opened up exciting and uncharted territories for humanities and social sciences research. In late medieval Europe, the production of charters drastically increased, resulting in the restriction of the scope of traditional 'diplomatics' research efforts to specific regions and collections. Enhancing conventional methodologies through the incorporation of digital methods came to be known as 'digital diplomatics'. Even with great potential, the digital diplomatics field has not yet reached its full capacity due to the scarcity of well-suited toolkits.

The analysis of textual patterns in documentary practices reaps considerable benefits from competent information search, retrieval, and exploration capabilities. In this context, information extraction and more precisely its named entity identification and classification subdomain, is considered among the first and most pertinent natural language processing steps (Ehrmann et al., 2023). Yet, the development of high-performing robust named entity recognition (NER) systems for historical texts is not straightforward and still faces substantial unresolved issues, namely input noisiness; language change; language diversity; document type and domain variety; and lack of resources.

This research endeavor is focused on medieval charters which are documents that record and formally express a legally enforceable act. Named entity recognition of medieval texts has been the focal point of several studies (Aguilar et al., 2016; Aguilar and Stutzmann, 2021; Chastang et al., 2021; Aguilar, 2022; Monroc et al., 2022) which collectively aimed at testing

†Corresponding Author

the applicability of contemporary NER techniques, such as feature-based machine learning methods and state-of-the-art deep learning-based ones, to the extraction of entities from charters written in a few medieval languages. A restrictive set of primarily five annotated corpora, widely shared among multiple studies, were employed in the training and evaluation of methods phases. Such top-notch manually curated datasets stand in stark contrast with real-world scenarios, where the prevalent issue of noisy input derived from handwritten text recognition errors leads to a notable drop in NER quality (Boroş et al., 2020).

To tackle this discrepancy, we propose investigating entity extraction from charters abstracts, which exhibit various advantageous features. Charters abstracts are brief summaries of the charters legal content which are written by expert historians in more modern languages. They are usually more available in digital archives than charters images, and contain the most prominent and 'diplomatics search relevant' named entities, that is the issuer name; the recipient name; and the place name (e.g., piece of land) which is often the object of exchange. That being said, they are commonly written in slightly outdated forms of present-day languages and might incorporate citations from the charters' original texts. The extraction of place names from Swedish medieval charters abstracts has been explored by (Karsvall and Borin, 2018), however they make use of a quite diminutive corpus that is composed of only 14 annotated charters abstracts, and on top of that they adopt a rule-based (modern Swedish) NER system.

For our starting trial, we seek to gain insight into how well an open-source Standard German NER system could function in this setting. To this end, we intend to assess the performance of a readily accessible pre-trained Standard German model against that of a custom-made NER model.

To build our ground-truth dataset, we randomly selected 2394 German abstracts of 138675 tokens from the archive of the Archbishopric of Salzburg and the Melk Abbey records which are publicly available on the monasterium.net platform. The tagging of person, place, and organization names in the abstracts was carried out following the classic BIO format which entails the usage of the B-TAG, I-TAG, and O labels to denote the beginning, continuation, and absence of named entities, respectively. The dataset was later partitioned into a training, validation, and test set, accounting for 70%, 15%, and 15% of the entire dataset, respectively.

We exploit the spaCy training functionality to train our NER model, and we select the spaCy large German pre-trained model (name: de_core_news_lg; architecture: CNN) as our Standard German NER model for performance comparison. We evaluate both systems on the test set using the typical token classification evaluation metrics: precision; recall; and

Experiment code is available on GitHub: https://github.com/Didip-eu/mom_ner

F1-Score. The obtained results, which are shown in Table 1, indicate that our trained NER system consistently outperforms the Standard German system by a noteworthy margin. Considering the relatively limited size of our training dataset, it is reasonable to infer that charters abstracts, even when scripted in modern languages, are likely to be too domain-specific for straightforward processing by regular NER systems.

Models / Categories	Trained NER Model			Standard German Model		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
B-PERS	0.89	0.85	**0.87**	0.62	0.56	0.59
I-PERS	0.92	0.82	**0.87**	0.89	0.48	0.62
B-LOC	0.73	0.53	**0.61**	0.34	0.69	0.46
I-LOC	0.37	0.11	**0.17**	0.14	0.20	**0.17**
B-ORG	0.81	0.65	**0.72**	0.11	0.05	0.07
I-ORG	0.83	0.67	**0.74**	0.21	0.05	0.07

Table 1: Evaluation Results on Test Set for the Trained NER Model and the Standard German Model

Future research could shift focus towards refining information extraction from medieval charters abstracts through the creation of finer resources (i.e, bigger datasets and more sophisticated software) and by targeting fine-grained NER.

Acknowledgements

This work is supported by the European Research Council advanced grant "From Digital to Distant Diplomatics" (No. 101019327).

References

Sergio Torres Aguilar. 2022. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the second workshop on language technologies for historical and ancient languages*, pages 119–128.

Sergio Torres Aguilar and Dominique Stutzmann. 2021. Named entity recognition for french medieval charters. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 37–46.

Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. 2016. Named entity recognition applied

on a data base of medieval latin charters. the case of chartae burgundiae. In *3rd International Workshop on Computational History (HistoInformatics 2016)*.

Emanuela Boroş, Verónica Romero, Martin Maarand, Kateřina Zenklová, Jitka Křečková, Enrique Vidal, Dominique Stutzmann, and Christopher Kermorvant. 2020. A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *2020 17th International conference on frontiers in handwriting recognition (ICFHR)*, pages 79–84. IEEE.

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

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.

Olof Karsvall and Lars Borin. 2018. Sdhk meets ner: Linking place names with medieval charters and historical maps. In *DHN*, pages 38–50.

Claire Bizon Monroc, Blanche Miret, Marie-Laurence Bonhomme, and Christopher Kermorvant. 2022. A comprehensive study of open-source libraries for named entity recognition on handwritten historical documents. In *International Workshop on Document Analysis Systems*, pages 429–444. Springer.