# Information Extraction from German Medieval Charters Abstracts

Sandy Aoun, Varvara Arzt, Daniel Luger, Georg Vogeler




# Diplomatics Research of Late Medieval Charters

**Historical Charters**:

- Documents that record and formally express a legally enforceable act
- Reflect the evolving relationships between rulers and subjects, as well as between different social and economic groups
- Crucial to understanding the development of legal and political systems throughout history

# Diplomatics Research of Late Medieval Charters

**Diplomatics:**

- Auxiliary sciences of history scholarly field
- Concerned with carrying in-depth critical analysis of historical documents
    - Understanding documentary practices:
        - Conventional visual features
        - Writing protocols and used textual patterns/formulae
- Applications:
    - Determining the authenticity of charters
    - Relating the information which the charters present to previously known facts

# Diplomatics Research of Late Medieval Charters

**Traditional Diplomatics** focuses on a limited number of charters at a time:

- Individual chanceries
- Collections of single institutions
- Practices of one country or region

**Surge in documentation of legal activity during the late medieval period**

**Monasterium.net Platform [1]**

| Period | Nbr of charters |
|---|---|
| Before 1300 | 66,156 |
| **14th century** | **156,098** |
| **15th century** | **164,376** |
| 16th century | 81,751 |
| 17th century | 56,271 |
| Not Dated | 132,898 |

# Diplomatics Research of Late Medieval Charters

Need for **‘Digital Diplomatics’**:

- Making use of the large amount of digitized historical records
- Incorporating digital methods into traditional diplomatics methodologies

Harnessing **digital technologies** to:

- Enable the study of large amounts of charters produced in different regions and/or institutions
- Open up the possibility of capturing a larger perspective on documentary practices

# Information Extraction from Late Medieval Charters

**Information Extraction**:

- Facilitates the semantic analysis of text
- Allows for automatic exploration and querying of massive datasets in a short period
- First relevant step: **Named Entity Recognition** (NER)

# NER of Medieval Charters

- Several studies: Aguilar et al., 2016; Chastang et al., 2021; Aguilar et al., 2021; Aguilar 2022; Monroc et al., 2022
- Testing the applicability of contemporary NER techniques to few medieval languages:
  - Feature-based Machine Learning methods vs. Deep Learning-based methods
- Restrictive set of annotated corpora to train and evaluate models (shared among multiple studies):
  - High-quality manually compiled datasets
  - Doesn’t account for the realistic case of noisy data due to HTR errors (Boroş et al., 2020)

# Proposal: Investigate NER of Charters Abstracts

**Charter Abstract:**

- Brief summary of a charter legal content
- Written by expert historian in more modern language

**Advantages:**

- More accessible in digital archives than charters images
- Contain important ‘diplomatics search relevant’ entities


“Receipt from Oswald Törringer, captain and steward of Mühldorf, for himself and others, that they received their pay for the campaign against Bohemia from the archbishopric.”

# NER of Charters Abstracts

Original Abstract:

> “**Christian Lanthaler (Länthaler), Bürger zu St. Veit im Pongau, verkauft seine Behausung und Hofstatt zu St. Veit dem Leonhard Ratzenberger.**”

EN Translation:

> “Christian Lanthaler (Länthaler), citizen of St. Veit im Pongau, sells his house and farmstead in St. Veit to Leonhard Ratzenberger.”

Abstract could:

- Include quote from text of original charter
- Be written in slightly outdated language form

# Experiment

Starting Point:

- **How well could an open-source Standard German NER system function in this setting?**

Experiment:

- Evaluate the performance of a **pre-trained Standard German model** against that of a **custom-made NER model**

# Building Ground-Truth Dataset

**Ground-Truth Dataset:**

- 2394 German abstracts of **138675 tokens** (randomly selected)
- From archive of the Archbishopric of Salzburg [2] and Melk Abbey records [3]
- Tagged NE types: **person, place, and organization** names
- Tagging scheme: **BIO format**
- Partitions: **70% training** set, **15% validation** set, **15% testing** set

# Custom and Pre-trained NER Model

**Trained NER Model:**

- Using training and validation sets
- Taking advantage of the spaCy training functionality to train model

**Pre-trained Standard German Model:**

- spaCy large German pre-trained model [4]
- Architecture: CNN

# Evaluation Results

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

# Perspectives

Since our training data is relatively small:

- Results imply that charters abstracts are too domain-specific to be handled easily by regular NER systems

Future Research:

- Creating bigger charters abstracts dataset
- Testing more cutting-edge models (BiLSTM-CRF, Embeddings)
- Exploring fine-grained NER

# References

**[1]** https://www.monasterium.net/mom/home

**[2]** https://www.monasterium.net/mom/AT-HHStA/SbgE/fond

**[3]** https://www.monasterium.net/mom/AT-StiAM/MelkOSB/fond

**[4]** https://spacy.io/models/de#de_core_news_lg

**Aguilar, S.T., Tannier, X. and Chastang, P., 2016.** Named entity recognition applied on a database of medieval latin charters. the case of chartae burgundiae. In 3rd International Workshop on Computational History (HistoInformatics 2016).

**Chastang, P., Aguilar, S.T. and Tannier, X., 2021.** A Named Entity Recognition Model for Medieval Latin Charters. Digital Humanities Quarterly, 15(4).

**Aguilar, S.T. and Stutzmann, D., 2021, December.** Named entity recognition for French medieval charters. In Proceedings of the Workshop on Natural Language Processing for Digital Humanities (pp. 37-46).

# References

**Aguilar, S.T., 2022, June.** 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 (pp. 119-128).

**Monroc, C.B., Miret, B., Bonhomme, M.L. and Kermorvant, C., 2022, May.** A comprehensive study of open-source libraries for named entity recognition on handwritten historical documents. In International Workshop on Document Analysis Systems (pp. 429-444). Cham: Springer International Publishing.

**Boroş, E., Romero, V., Maarand, M., Zenklová, K., Křečková, J., Vidal, E., Stutzmann, D. and Kermorvant, C., 2020, September.** 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) (pp. 79-84). IEEE.

# Thank You for Listening!!