People and Places of Historical Europe: Bootstrapping Annotation Pipeline and a New Corpus of Named Entities in Late Medieval Texts

Vít Novotný[1] and **Kristýna Luger**[2] and **Michal Štefánik**[1]
Tereza Vrabcová[1] and **Aleš Horák**[1]

[1]Faculty of Informatics, Masaryk University, Brno, Czech Republic
[2]Faculty of Arts, Masaryk University, Brno, Czech Republic
{witiko,449852,stefanik.m,485431,hales}@mail.muni.cz

Abstract

Although pre-trained named entity recognition (NER) models are highly accurate on modern corpora, they underperform on historical texts due to differences in language OCR errors. In this work, we develop a new NER corpus of 3.6M sentences from late medieval charters written mainly in Czech, Latin, and German.

We show that we can start with a list of known historical figures and locations and an unannotated corpus of historical texts, and use information retrieval techniques to automatically bootstrap a NER-annotated corpus. Using our corpus, we train a NER model that achieves entity-level Precision of 72.81–93.98% with 58.14–81.77% Recall on a manually-annotated test dataset. Furthermore, we show that using a weighted loss function helps to combat class imbalance in token classification tasks. To make it easy for others to reproduce and build upon our work, we publicly release our corpus, models, and experimental code.

1 Introduction

Named entity recognition (NER) techniques enable the extraction of valuable insights from unstructured information in various domains. With the advancements in optical character recognition (OCR, Breuel, 2017; Kodym and Hradiš, 2021), NER can now be applied to scanned historical texts spanning over a millennium. However, despite the significant interest in NER, resources for training models to recognize entities in medieval texts remain scarce (Ehrmann et al., 2021, Table 3).

In this work, we present a new multilingual NER corpus of 3.6M sentences from the AHISTO project, which aims to build a searchable web database of late medieval charters from scanned images. In Section 2, we review recent related work in historical NER. In Section 3 and 4, we describe the database and the automatic pipeline used to bootstrap our corpus in. In Section 5, we use our corpus to train models for historical NER and evaluate them on a manually-annotated test dataset. We show that our models are highly-accurate, reaching Precision of up to 94% with up to 82% recall. Additionally, we show that the use of a weighted loss function is crucial in token classification tasks with high class imbalance. We conclude in Section 7.

To facilitate reproducibility and future research, we publicly release our corpus under open CC0 license as well as our models and code.[1]

2 Related Work

Grover et al. (2008) created a corpus of British parliamentary proceedings from the late 17th and early 19th centuries using OCR techniques and expert annotation. They developed and evaluated a rule-based NER classifier, achieving an F_1-score of 71%, noting the detrimental effect of OCR errors.

Hubková et al. (2020) created a corpus of 32 historical Czech newspapers from 1872 using OCR and expert annotation. Hubková and Král (2021) achieved a state-of-the-art F_1-score of 82% on the corpus by fine-tuning a pre-trained SlavicBERT model (Arkhipov et al., 2019).

Blouin et al. (2021) fine-tuned pre-trained monolingual Transformer models on early modern NER corpora in English, French, and German, achieving a near-state-of-the-art F_1-score of 62%. They also found that the character-based CharBERT model (Ma et al., 2020) was more robust against OCR errors than BERT (Devlin et al., 2019).

Torres Aguilar (2022) developed a corpus of 7,576 medieval charters ranging from the 10th century to the 15th century using expert annotation. They showed that finetuning pre-trained multilingual Transformer models such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) gave comparable results to state-of-the-art Bi-LSTM-CRF NER models (Ma and Hovy, 2016).

[1]https://nlp.fi.muni.cz/projects/ahisto/ner-resources

Findings of the Association for Computational Linguistics: ACL 2023, pages 14104–14113

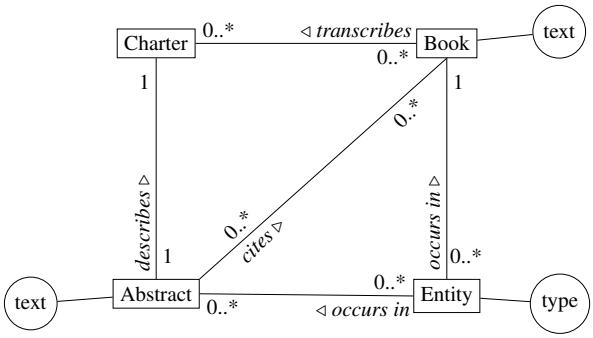


Figure 1: Entity relationship diagram of the database that was produced in the AHISTO project.

3 Data Description

The AHISTO project database includes various document types, as shown in Figure 1. The main focus is on *charters* from Europe during the Hussite era (1419–1436), each with an accompanying *abstract* written by project historians in contemporary Czech. Historians also manually annotate all named entities (people and places) in the abstracts. The original text of charters, if available, can be found in *books*. OCR texts of books are available (Novotný et al., 2021; Novotný and Horák, 2022), but named entities are not annotated.

The database includes 4,182 abstracts with 5,621 sentences, and 15,100 unique named entities: people (62.53%) and places (37.47%). It also includes 872 books with 268,669 pages and 3.6M sentences, mostly in medieval Czech (43.23%), Latin (36.32%), and German (16.89%). Of these pages, 3,553 (1.32%) with 50k sentences were selected as relevant to medieval charters by project historians.

4 Corpus Annotation

We developed five NER corpora from the database of the AHISTO project. We created **Abstracts-Tiny** from abstracts and **Books-Small**, **Medium**, **Large**, and **Huge** from books. The statistics of all our corpora are listed in Table 1.

4.1 Corpus from Abstracts

For the evaluation of NER models on contemporary Czech texts that discuss medieval charters, we constructed a corpus **Abstracts-Tiny** from all 5,621 sentences in abstracts. We randomly split the sentences into 80%, 10%, and 10% for training, validation, and testing.

4.2 Bootstrapping Initial Corpus from Books

To bootstrap our initial corpus **Books-Small**, we used the information retrieval system from the Manatee library (Rychlý, 2007; Bušta et al., 2023), which performed the best out of 9 systems that we considered, see also Appendix A. First, we indexed OCR texts from the 3,553 book pages that historians selected as relevant. Then, for each of the 15,100 named entities in abstracts, we used a boolean phrase query to retrieve all occurrences of the named entity in the index. For each occurrence of a named entity, we extracted the surrounding sentence and we merged all sentences that were extracted multiple times.

Table 1: The numbers of sentences and the occurrences of people (B-PER tokens) and places (B-LOC tokens) in our NER corpora. For each corpus, we report statistics for training, validation, and testing splits.

Corpus	# Sentences	# B-PER	# B-LOC
Abstracts-Tiny	**5,222**	**10,981**	**5,933**
Training	4,320	9,032	4,952
Validation	502	1,160	606
Testing	400	789	375
Books-Small	**7,842**	**4,778**	**5,679**
Training	6,493	3,877	4,722
Validation	1,249	614	714
Testing	100	287	243
Books-Medium	**7,842**	**17,400**	**17,184**
Training	6,493	13,958	13,987
Validation	1,249	3,155	2,954
Testing	100	287	243
Books-Large	**46,739**	**46,051**	**45,435**
Training	44,155	43,360	42,315
Validation	2,484	2,404	2,877
Testing	100	287	243
Books-Huge	**3,629,903**	**4,257,380**	**2,865,470**
Training	3,227,624	3,794,991	2,545,820
Validation	402,179	462,102	319,407
Testing	100	287	243

We randomly split the sentences into 80%, 10%, and 10% for training, validation, and testing of NER models. From the testing split, we randomly sampled 100 sentences and a volunteer Czech graduate student of history manually checked all entities in the sentences. See Appendix B for annotator instructions. We used the 100 sentences for testing.

4.3 Inferring Missing Entities in Books

Sentences in the **Books-Small** corpus contained many named entities that were not part of abstracts and were therefore missing. To produce our intermediate **Books-Medium** corpus, we trained a NER model on the **Books-Small** corpus and used it to infer the missing named entities in **Books-Small**.

Most named entities in the **Books-Medium** corpus were annotated, but the corpus only contained a small portion of the 3.6M sentences in all books. To produce our final **Books-Large** corpus, we trained

a NER model on the **Books-Medium** corpus and used it to infer named entities in all book pages that historians selected as relevant. Furthermore, we also inferred named entities in all books to produce the corpus **Books-Huge**, which is $100\times$ larger than **Books-Large** but may contain irrelevant sentences.

We describe the NER models that we used for the inference in the following section.

5 Experiments

In this section, we describe the NER models that we used to produce the **Books-Medium**, **Large**, and **Huge** corpora and how we evaluated them.

5.1 Models

We trained two models by fine-tuning pretrained XLM-RoBERTa models (Conneau et al., 2020) using the Adaptor library (Štefánik et al., 2022) for multi-objective training.

To produce the **Books-Medium** corpus, we fine-tuned the XLM-RoBERTa-Base model (125M parameters). To produce the **Books-Large** and **Books-Huge** corpora, we fine-tuned the XLM-RoBERTa-Large model (355M parameters). To simplify the discussion, we will refer to the models as **Model S** (for small) and **Model L** (for large) throughout the paper. See Appendix C for the description of our hardware and hyperparameters.

Given the scarcity of pre-trained historical NER models that are publicly available, we compare our model against the XLM-RoBERTa-Large model fine-tuned on contemporary German news texts (Gugger and Gerchick, 2022) from the CoNLL03 dataset (Tjong Kim Sang and De Meulder, 2003). Although the model was only trained on German data, Ruder et al. (2019) show that the model should generalize well to other languages.

5.2 Objectives

In order to train our models effectively, we used a multi-objective approach, utilizing two distinct objectives in our optimization process:

Masked Language Modeling (MLM): Unsupervised regression on the **Books-Large** corpus.

Token Classification (TC): Supervised classification of tokens into classes B-PER, I-PER, B-LOC, I-LOC, and O. To address the issue of class imbalance, we use the weighted cross-entropy (WCE) loss function with inverse class frequencies as weights.

We adopt a sequential schedule for alternating between objectives, where each objective is trained for a single epoch. This approach has the advantage of allowing focused optimization of each objective.

5.3 Quantitative Evaluation

We use the micro-averaged token-level F_β-score with a value of $\beta = 0.25$ as the evaluation metric for both the validation of the TC objective and the quantitative evaluation of our models. This metric prioritizes precision in entity recognition, even if it results in lower recall. To evaluate out-of-domain performance on contemporary Czech texts that discuss medieval charters, we report the F_β-score on two benchmarks: the **Abstracts-Tiny** corpus and the **Books-*** corpora.

In addition to the token-level F_β-score, we also present the entity-level Precision and Recall. These measures are considered to be more representative for the majority of NER applications, as opposed to per-token evaluation measures, which are tied to the tokenizer of a model. Similarly to Ehrmann et al. (2020), we use two evaluation regimes:

Strict Predicted entities must match both the type and boundaries of expected entities.

Fuzzy Predicted entities must match the type and overlap the boundaries of any expected entity.

We report the micro-averaged Precision and Recall for our best model as a range of Strict–Fuzzy%. We report both overall and per-language results.

5.4 Qualitative Evaluation

In order to conduct a comprehensive qualitative evaluation, we also report a confusion matrix of our best model on the **Books-*** corpora. In Appendix D, we also compare the predictions made by our best model with manual annotations.

5.5 Ablation Study

In this section, we present a series of ablation experiments designed to investigate the impact of using various training data and loss functions for the TC objective. In the experiments, we fine-tune the small XLM-RoBERTa-Base model due to environmental considerations.

Training Data Size We evaluate the validity of our annotations by training the TC objective not only on the **Books-Medium** corpus, but also the **Books-Small** and **Large** corpora. We will refer to the models as **Model TDS1** and **Model TDS2**.

Table 2: Evaluation results for our NER models and the baseline. For each model, we list a short identifier for ease of reference, the size of the model in parameters, the training data and the loss function of the Token Classification (TC) objective, the loss function used in the TC objective, and the accuracy measured by per-token F_β-score on both the **Abstracts-Tiny** corpus and the **Books-*** corpora. Best results are **bold**.

Model Id	Training Data	Model Size	TC Loss	F_β-score Abstracts-Tiny	F_β-score Books-*
L	Books-Medium	355M	WCE	**91.61%**	**93.43%**
S	Books-Medium	125M	WCE	90.51%	93.19%
TDS2	Books-Large	125M	WCE	87.21%	89.11%
TDS1	Books-Small	125M	WCE	88.92%	88.20%
CI	Books-Small	125M	CE	86.41%	84.45%
	CoNLL03 (de)	355M	CE	80.59%	80.74%

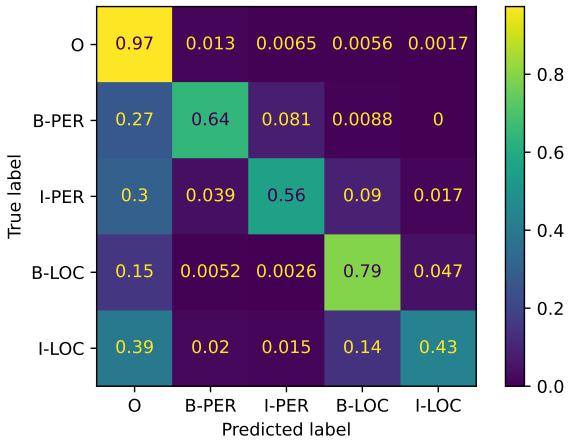


Figure 2: A confusion matrix of our model L on the **Books-*** corpora. Best viewed in color.

Class Imbalance We replaced the weighted cross-entropy loss function in the TC objective with unweighted cross-entropy loss in the **Books-Small** corpus to investigate the impact of missing named entities. We will refer to this model as **Model CI**.

6 Discussion

6.1 Quantitative Evaluation

Table 2 shows that both **Model L** and **Model S** achieved a per-token F_β-score of over 90% on both benchmarks, outperforming the baseline model by more than 10% on both benchmarks.

Despite its smaller size, **Model S** received only 1.1% less per-token F_β-score than **Model L** on the **Abstracts-Tiny** corpus and only 0.24% less on the **Books-*** corpora. This makes **Model S** a compelling choice for low-resource NER applications.

The per-entity Precision of **Model L** on the **Books-*** corpora was 72.81–93.98% with 58.14–81.77% Recall. This shows that our model can reliably recognize named entities, even though it does not always exactly match the boundaries. Per-language, Precision was the highest for Czech (77.42–95.63%) and the lowest for German (70.80–87.07%).

6.2 Qualitative Evaluation

Figure 2 reveals that **Model L** is more accurate at identifying the beginnings of named entities than their endings. This is evidenced by the smaller probability of misclassifying tokens B-PER and B-LOC as class O ($\leq 21\%$) compared to the I-PER and I-LOC tokens ($\geq 28\%$).

Our model will also sometimes recognize a list of places as a single named entity or divide a place name into several named entities. This is evidenced by the relatively high probability of classifying tokens B-LOC as I-LOC (16%) and vice versa (20%).

Our model will seldom make a mistake in identifying the type of a named entity. This is evidenced by the low probability of misclassifying tokens *-PER as *-LOC ($\leq 1\%$) and vice versa ($\leq 2.4\%$).

6.3 Ablation Study

Table 2 shows that both **Model TDS1** and **Model TDS2** performed worse than **Model L** on both benchmarks. **Model TDS1** was affected by the class imbalance in the **Books-Small** corpus, missing many named entities. **Model TDS2** suggests that even though the **Books-Large** corpus was constructed from pages relevant to medieval charters, there may be irrelevant sentences within the pages.

Model CI received 2.51% less per-token F_β-score than **Model TDS1** on the **Abstracts-Tiny** corpus and 3.75% less on the **Books-*** corpora. This shows that using weighted loss is crucial in token classification tasks with high class imbalance.

7 Conclusion

Despite the large interest in named entity recognition (NER) in the last few decades, studies targeting late medieval historical texts are still scarce.

In our work, we have developed a new silver-standard NER corpus of 3.6M sentences from late medieval charters. We described our automatic pipeline for bootstrapping a corpus using a list of known named entities. We also showed that our corpus can be used to train highly accurate models for historical NER. Lastly, we have demonstrated the usefulness of using weighted loss functions in token classification tasks with high class imbalance.

8 Limitations

In this study, we assessed the quality of the corpora **Books-Small**, **Medium**, and **Large**, by training and evaluating a NER model on them but we did not include the corpus **Books-Huge** in our analysis. However, our results on the **Books-Large** corpus indicate that there is no substantial benefit to using a corpus larger than **Books-Medium** for training NER models. This is consistent with prior research on few-shot training on smaller corpora achieving comparable accuracy to larger, potentially noisy corpora (Blouin et al., 2021).

Given the scarcity of benchmarks for late medieval NER (Ehrmann et al., 2021, Table 3), we were unable to conduct experiments on corpora other than our own. Additionally, we utilized a NER model trained on contemporary texts as our baseline for comparison. Therefore, it is important to note that these results may not generalize to other medieval NER tasks. In the future, efforts should be made to develop more comprehensive benchmarks for late medieval NER such as our own.

9 Ethical Considerations

In conducting this research, we were committed to upholding the ethical principles of respect for persons, beneficence, and non-maleficence. Specifically, the annotation in this work was carried out voluntarily by informed participants, and the welfare and rights of these participants were protected throughout the research process. Furthermore, the annotation did not involve collecting any personal or sensitive information from individuals.

Acknowledgements

This work has been produced with the assistance of the Czech Medieval Sources online database, provided by the LINDAT/CLARIAH-CZ research infrastructure, supported by the Ministry of Education, Youth, and Sports of the Czech Republic (Project No. LM2018101).

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Baptiste Blouin, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. Transferring modern named entity recognition to the historical domain: How to take the step? In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 152–162, NIT Silchar, India. NLP Association of India (NLPAI).

Thomas M Breuel. 2017. High performance text recognition using a hybrid convolutional-LSTM implementation. In *14th IAPR international conference on document analysis and recognition (ICDAR 2017)*, volume 1, pages 11–16. IEEE.

Jan Bušta, Miloš Jakubíček, Michal Cukr, Ondřej Herman, and Marek Medveď. 2023. Nosketch engine.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451. ACL.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *SIGIR*, pages 758–759, New York, NY, USA. ACL.

Dong Deng, Guoliang Li, Jianhua Feng, Yi Duan, and Zhiguo Gong. 2015. A unified framework for approximate dictionary-based entity extraction. *The VLDB Journal*, 24(1):143–167.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. This paper is a preprint and has not been peer-reviewed.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310, Cham. Springer.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Sylvain Gugger and Marissa Gerchick. 2022. xlm-roberta-large-finetuned-conll03-german.

Helena Hubková and Pavel Král. 2021. Transfer learning for Czech historical named entity recognition. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 576–582, Held Online. INCOMA Ltd.

Helena Hubková, Pavel Král, and Eva Pettersson. 2020. Czech historical named entity corpus v1.0. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4458–4465.

Oldřich Kodym and Michal Hradiš. 2021. Page layout analysis system for unconstrained historic documents. In *International Conference on Document Analysis and Recognition, ICDAR 2021*, pages 492–506. Springer.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Gonzalo Navarro. 2003. Approximate regular expression searching with arbitrary integer weights. In *Algorithms and Computation*, pages 230–239, Berlin, Heidelberg. Springer.

Vít Novotný and Aleš Horák. 2022. When tesseract meets PERO: Open-source optical character recognition of medieval texts. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022*, pages 157–161. Tribun EU.

Vít Novotný, Kristýna Seidlová, Tereza Vrabcová, and Aleš Horák. 2021. When tesseract brings friends: Layout analysis, language identification, and super-resolution in the optical character recognition of medieval texts. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021*, pages 29–39. Tribun EU.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP 2019*, pages 3982–3992, Hong Kong, China. ACL.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *NIST Special Publication*, pages 109–126.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Pavel Rychlý. 2007. Manatee/Bonito: A modular corpus manager. In *RASLAN*, pages 65–70. Tribun EU.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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, Marseille, France. European Language Resources Association.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *ICLR*.

Michal Štefánik, Vít Novotný, Nikola Groverová, and Petr Sojka. 2022. Adaptor: Objective-centric adaptation framework for language models. In *ACL*, pages 261–269, Dublin, Ireland. ACL.

A Comparison of Bootstrapping Methods

For all the 15,100 named entities in abstracts, we used several information retrieval techniques to find their occurrences in the books.

First, we used four fast techniques to produce lists of up to 10,000 candidate results:

1. **Jaccard Similarity** (Deng et al., 2015): From the books, we extracted substrings of length that were similar to the length of the current entity. We ordered the substrings by their character and word Jaccard similarity to the current entity.

2. **Okapi BM25** (Robertson et al., 1995): From the books, we extracted phrases of length that were similar to the length of the current entity and we indexed them as individual retrieval units in an inverted index. We retrieved the phrases using a ranked retrieval query for the current entity, using BM25 weighting.

3. **Manatee** (Rychlý, 2007; Bušta et al., 2023): From the books, we extracted lemmatized tokens and we indexed them as individual retrieval units in a positional inverted index. We retrieved phrases using a boolean phrase query for the current entity. Since boolean retrieval results are not ranked, we ranked them by their character edit distance to the current entity.

Table 3: The Precision, Recall, and F_β-score ($\beta = 0.25$) for the different retrieval techniques that we tried for bootstrapping our **Books-Small** corpus,

Method	Precision	Recall	F_β-score
Manatee	**100.00%**	17.34%	**78.10%**
Fuzzy Regexes	78.98%	23.40%	69.30%
Edit Distance	74.00%	**24.92%**	66.32%
Concatenation	72.50%	24.41%	64.97%
BERTScore	70.50%	23.74%	63.18%
SentenceBERT	69.50%	23.40%	62.28%
Jaccard Similarity	63.00%	21.21%	56.46%
Reciprocal Rank Fusion	62.00%	20.88%	55.56%
Okapi BM25	35.03%	11.62%	31.31%

4. **Fuzzy Regexes** (Navarro, 2003): From the current entity, we extracted an approximate regular expression. From the books, we retrieved all substrings that matched the regular expression up to a certain edit distance.

Then, we used three slow techniques to rerank all candidate results for each named entity:

1. **Edit Distance**: We reordered the results by their word and character edit distance to the current entity.

2. **BERTScore** (Zhang et al., 2020): We reordered the results by their BERT F_1-score to the current entity.

3. **SentenceBERT** (Reimers and Gurevych, 2019): We reordered the results by the cosine similarity of their SentenceBERT embeddings to the embedding of the current entity.

Finally, we used two rank fusion techniques to combine the results of the above techniques:

1. **Reciprocal Rank Fusion** (Cormack et al., 2009): We combined the results of all the inexpensive and expensive techniques based on the ranks of the results across the techniques.

2. **Concatenation**: We started with the results produced by Fuzzy Regexes, if any, followed by the results produced by the Reciprocal Rank Fusion. For duplicate results, we kept the results from Fuzzy Regexes.

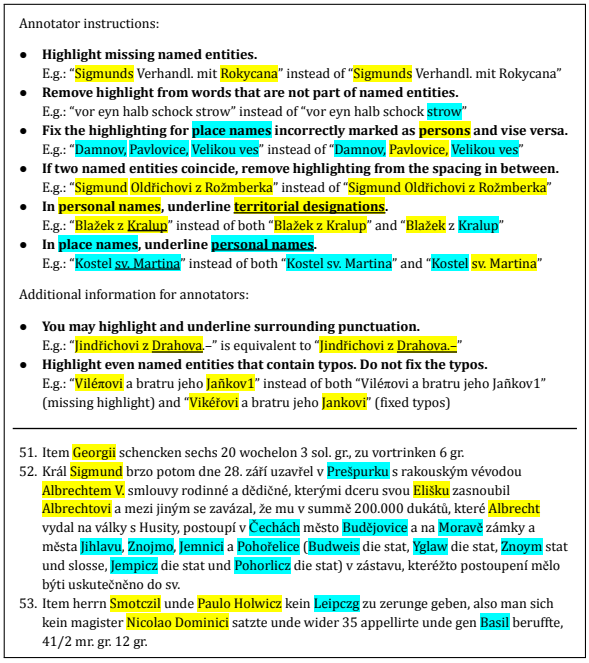
To select the best retrieval technique, we sampled 21 named entities from the abstracts and used each of the methods to produce up to 10 results. Then, three Czech experts employed as investigators in the AHISTO project annotated the relevance of the results. Using the annotations, we computed Precision, Recall, and F_β-score ($\beta = 0.25$), see Table 3.

Based on F_β-score, we selected Manatee as the the best technique. The high accuracy of Manatee and Fuzzy Regexes shows that lemmatization and approximate search are important for the retrieval of named entities in OCR texts because they help with morphological variations and OCR errors.

B Annotator Instructions

Figure 3 shows the interface for the collection of manual annotations for the **Books-*** corpora.

Annotators were instructed to identify nested named entities, including territorial designations (e.g. Blažek of *Kralupy*) and dedications of buildings (e.g. Church of *St. Martin*). These nested annotations were utilized in the evaluation of NER models to prevent penalization for recognizing nested named entities as separate entities.


Annotator instructions:

- **Highlight missing named entities.**
  E.g.: "Sigmunds Verhandl. mit Rokycana" instead of "Sigmunds Verhandl. mit Rokycana"
- **Remove highlight from words that are not part of named entities.**
  E.g.: "vor eyn halb schock strow" instead of "vor eyn halb schock strow"
- **Fix the highlighting for place names incorrectly marked as persons and vise versa.**
  E.g.: "Damnov, Pavlovice, Velikou ves" instead of "Damnov, Pavlovice, Velikou ves"
- **If two named entities coincide, remove highlighting from the spacing in between.**
  E.g.: "Sigmund Oldřichovi z Rožmberka" instead of "Sigmund Oldřichovi z Rožmberka"
- **In personal names, underline territorial designations.**
  E.g.: "Blažek z Kralup" instead of both "Blažek z Kralup" and "Blažek z Kralup"
- **In place names, underline personal names.**
  E.g.: "Kostel sv. Martina" instead of both "Kostel sv. Martina" and "Kostel sv. Martina"

Additional information for annotators:

- **You may highlight and underline surrounding punctuation.**
  E.g.: "Jindřichovi z Drahova.–" is equivalent to "Jindřichovi z Drahova.–"
- **Highlight even named entities that contain typos. Do not fix the typos.**
  E.g.: "Viléřovi a bratru jeho Jaňkov1" instead of both "Viléřovi a bratru jeho Jaňkov1" (missing highlight) and "Vikéřovi a bratru jeho Jankovi" (fixed typos)

51. Item Georgii schencken sechs 20 wochelon 3 sol. gr., zu vortrinken 6 gr.
52. Král Sigmund brzo potom dne 28. září uzavřel v Přešpurku s rakouským vévodou Albrechtem V. smlouvy rodinné a dědičné, kterými dceru svou Elišku zasnoubil Albrechtovi a mezi jiným se zavázal, že mu v summě 200.000 dukátů, které Albrecht vydal na války s Husity, postoupí v Čechách město Budějovice a na Moravě zámky a města Jihlavu, Znojmo, Jemnici a Pohořelice (Budweis die stat, Yglaw die stat, Znoym stat und slosse, Jempicz die stat und Pohorlicz die stat) v zástavu, kteréžto postoupení mělo býti uskutečněno do sv.
53. Item herrn Smotczil unde Paulo Holwicz kein Leipczg zu zerunge geben, also man sich kein magister Nicolao Dominici satzte unde wider 35 appellirte unde gen Basil beruffte, 41/2 mr. gr. 12 gr.

Figure 3: The interface for the annotation of our test dataset using the Google Documents web service. The interface includes annotator instructions (top) and an ordered list of sentences from books (bottom).

C Training Details

To fine-tune the XLM-RoBERTa-Base model, we used a learning rate of $5 \cdot 10^{-5}$ with a linear decay until reaching 10 total training epochs or until convergence on the validation dataset. The fine-tuning took approximately 10 GPU hours on an NVIDIA Tesla T4 graphics card.

Table 4: Example sentences in different languages from the **Abstracts-Tiny** and **Books-*** corpora. Manual annotations are compared to the predictions of **Model L**. Person names are **bold** and place names are *in italics*. Nested named entities such as territorial designations and designations of buildings are ***both bold and italic***.

Corpus	Language	Sentence Manually annotated ground truth	Model prediction
Abstracts-Tiny	Czech	**Bohuněk a Kundrát, bratři z *Miroslavi,*** na základě svolení od **probošta *dolnokounického kláštera* Jana** a převorky, že mohou rozšířit rybník v *Hlavaticích,* slibují jen na jistou vzdálenost od *šumického dvora* zatopit a dovolit šumickým lidem, aby užívali tamní potok.	**Bohuněk** a **Kundrát,** bratři z *Miroslavi,* na základě svolení od probošta dolnokounického kláštera **Jana** a převorky, že mohou rozšířit rybník v *Hlavaticích,* slibují jen na jistou vzdálenost od šumického dvora zatopit a dovolit šumickým lidem, aby užívali tamní potok.
Books-*	Czech	Vedle mnohaletého tohoto hejtmana čáslavského je tu **Žižkův** bratr **Jaroslav,** známí nám **bratří Valečovští,** sirotčí pozdější hejtmané **Jíra z *Řečice*** (*Koudelova* u *Čáslavě*) a **Blažek z *Kralup*,** táborský **Jakub Kroměšín** a mnoho jiných statečných válečníků, i nejeden prostý voják, který však v **Žižkově** radě zasedá jako rovný s urozenými.	Vedle mnohaletého tohoto hejtmana čáslavského je tu **Žižkův** bratr **Jaroslav,** známí nám bratří **Valečovští,** sirotčí pozdější hejtmané **Jíra z Řečice** (*Koudelova* u *Čáslavě*) a **Blažek z Kralup,** táborský **Jakub Kroměšín** a mnoho jiných statečných válečníků, i nejeden prostý voják, který však v *Žižkově* radě zasedá jako rovný s urozenými.
Books-*	Latin	**Johannis Rupolth** vac., ad present. nobilis **Hinconis Berka de Duba** residentis in castro *Scharffstein.* **Exec. pleb. in *Arnorssdorff.*** C, IIII.- *Horzielicz.*- Anno quo supra die XXVI April. data e. crida **Thome,** clerico de *Antiqua Boleslauia,* ad eccl. paroch.	**Johannis Rupolth** vac., ad present. nobilis **Hinconis Berka de Duba residentis in castro Scharffstein.** Exec. pleb. in *Arnorssdorff.* C, IIII.- *Horzielicz.*- Anno quo supra die XXVI April. data e. crida **Thome,** clerico de *Antiqua Boleslauia,* ad eccl. paroch.
Books-*	German	September 3. Der Rat zu *Löbau* leiht 160 Schock zum Bau und zur Besserung der durch Brand und die Ketzer zerstörten Stadt. Nach Knothe Urkundenbuch von *Kamenz* und *Löbau* S. 253 (nach dem Original im *Löbaner Stadtarchiv,* jetzt im *Hauptstaatsarchiv* zu *Dresden).* 25 1432. September 12. Item **Nickel Windischs** ist ufgenomen des freitags vor des heiligen creucis exaltation.	September 3. Der Rat zu *Löbau* leiht 160 Schock zum Bau und zur Besserung der durch **Brand** und die Ketzer zerstörten Stadt. Nach Knothe Urkundenbuch von *Kamenz* und *Löbau* S. 253 (nach dem Original im *Löbaner Stadtarchiv,* jetzt im Hauptstaatsarchiv zu *Dresden).* 25 1432. September 12. **Item Nickel Windischs** ist ufgenomen des freitags vor des heiligen creucis exaltation.

To fine-tune the XLM-RoBERTa-Large model, we used a smaller learning rate of $5 \cdot 10^{-6}$ with a warm-up period of 20 epochs to mitigate overfitting and improve generalization performance. After the initial 20 epochs, we used a linear decay until reaching 200 total training epochs. The fine-tuning took approximately 74 GPU hours on an NVIDIA A40 graphics card.

The total computational budget of our project was approximately 114 GPU hours.

D Model Predictions

We randomly sampled three sentences from the testing split of the **Books-*** corpora, written in historical Czech, Latin, and German, as well as one sentence from the testing split of the **Abstracts-Tiny** corpus, written in modern Czech. Then, we compared the prediction of **Model L** with manual annotations for these four sentences.

Table 4 illustrates that while **Model L** achieved a high level of precision, it failed to recall at least one named entity in every example sentence except the Czech sentence from the **Books-*** corpora. Furthermore, our model occasionally misidentifies the boundaries of named entities. An example of this can be observed in the Czech sentence from the **Books-*** corpora, where the named entity "bratří Valečovští" (translated as "the Valečov brothers") is incorrectly shortened to simply "Valečovští" (translated as "the Valečovs"). We can see the opposite error in the Latin sentence, where the personal name "Hinconis Berka de Duba" was incorrectly extended to include Berka's place of residence.

Compared to failures to recall named entities and identify their boundaries, errors in detecting types of named entities are rare as we already showed in Section 6.2. The only example of an incorrectly predicted type occurs in the Czech sentence from the **Books-*** corpora, where the personal name "Žižkově" (translated as "Žižka's") was mistaken for Žižkov, a district of Prague, despite the presence of ample context clues in the surrounding sentence.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discuss the limitations of our study in Section 8 (Limitations).

☑ A2. Did you discuss any potential risks of your work?
We discuss the ethical considerations of our study in Section 9 (Ethical Considerations).

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The claims we make in the abstract and introduction are substantiated in Section 4 (Corpus Annotation) and 6 (Discussion).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We list the scientific artifacts we created in Section 4 (Corpus Annotation) and 5.1 (Models).

☑ B1. Did you cite the creators of artifacts you used?
We cite the scientific artifacts we used in Section 5.1 (Models). We do not disclose the source of the data in Section 3 (Data Description) to ensure anonymity for the review.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We disclose the license of the artifacts we created in Section 1 (Introduction).

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We disclose the terms of use of the artifacts we created in Section 1 (Introduction).

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. All personal information that was collected pertains to medieval historical figures. Therefore, the issue of anonymization was not a concern for us as the individuals in question are not living and do not have any personal information that can be potentially compromised.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We document the scientific artifacts we created in Sections 3 (Data Description) and 4 (Corpus Annotation).

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We report statistics and splits of the data we created in Section 4 (Corpus Annotation).

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

We describe our experiments in Section 5 (Experiments).

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We describe the number of parameters in the models used in Section 5.1 (Models). We describe our computing infrastructure and the total computational budged in Appendix C (Training Details).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss the experimental setup in Section 5 (Experiments). We describe the hyperparameters used in Appendix C (Training Details).

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We describe the details of the descriptive statistics we report in Section 5.3 (Quantitative Evaluation).

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We report that we used the AdaptOr library for multi-objective training in Section 5.1 (Models). We disclose the parameter settings used in Section 5.2 (Objectives).

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

We report the involvement of human annotators in Section 3 (Data Description) and 4 (Corpus Annotation) and in Appendix A (Comparison of Bootstrapping Methods).

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We report annotator instructions in Appendix B (Annotator Instructions).

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We discuss how participants were selected and compensated in Section 3 (Data Description) and 4 (Corpus Annotation) and in Appendix A (Comparison of Bootstrapping Methods).

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. All personal information that was collected pertains to medieval historical figures. Therefore, the issue of consent was not a concern for us as the individuals in question are not living and do not have any personal information that can be potentially compromised.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our annotation does not involve collecting any personal or sensitive information from individuals. As such, it does not raise any ethical concerns and does not require approval by an ethics review board.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We report the basic demographic and geographic characteristics of the annotator population in Section 3 (Data Description) and 4 (Corpus Annotation) and in Appendix A (Comparison of Bootstrapping Methods).