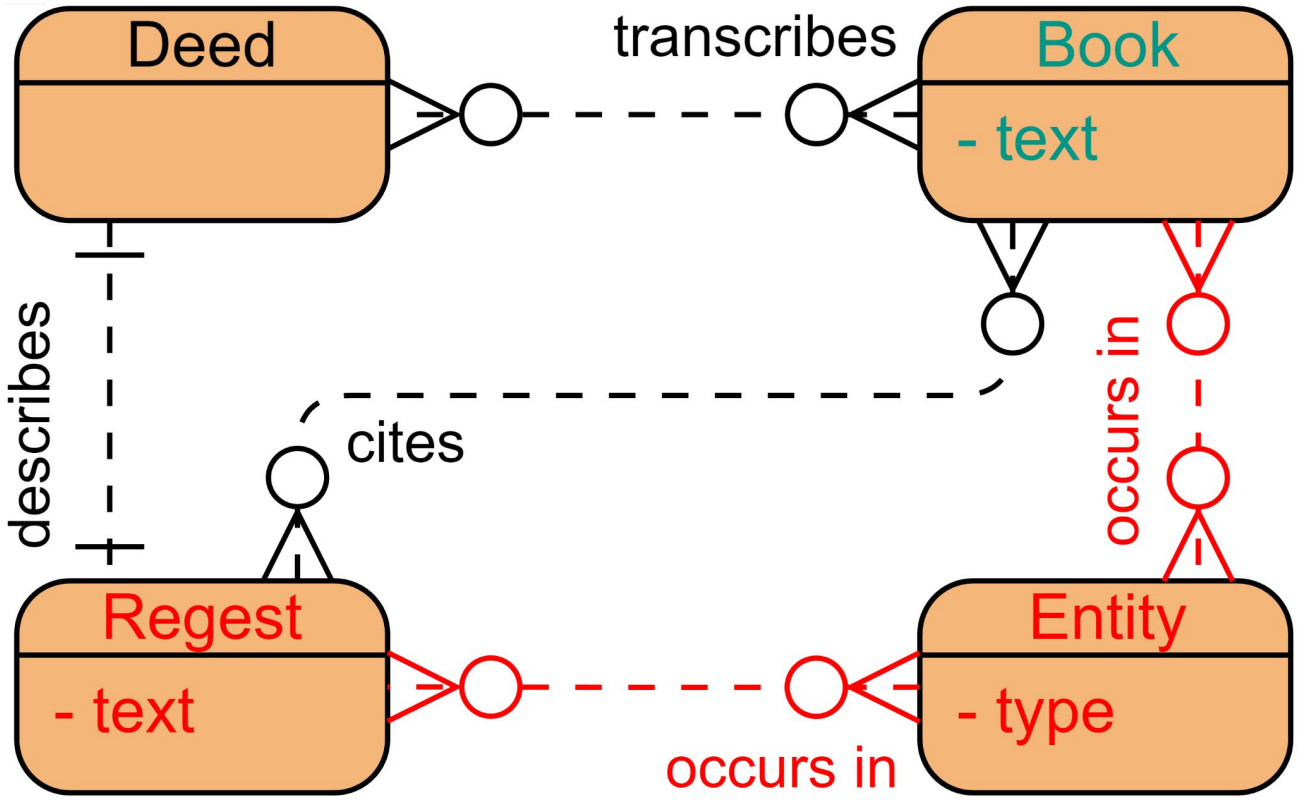
MUNI
FI

# The Retrieval and Recognition of Named Entities in Medieval Texts

Vít Novotný

# Introduction

## Entity-Relationship Diagram of AHISTO




- 814 *books* transcribe medieval *deeds*.
- During 2020–2021, we focused at recognizing text in scanned book *pages*:
  - NOVOTNÝ. When Tesseract Does It Alone. In Horák et al. *RASLAN 2020*.
  - NOVOTNÝ et al. When Tesseract Brings Friends. In Horák et al. *RASLAN 2021*.
- Meanwhile, our colleagues from ARTS wrote 2 094 *regests* that describe *deeds*, denote *entities* (people and places) that occur in deeds, and cite related *books*.
- In 2022, we focused at retrieving *entities* from *regests* in *book* texts and recognizing new *entities* in *book* texts.



MUNI FI

# Methods

## Named Entity Retrieval

- For retrieval, we use a number of inexpensive retrieval techniques:
  - Jaccard Similarity
  - Okapi BM25
  - Fuzzy regexes
  - Manatee
- For reranking, we use a number of expensive retrieval techniques:
  - Edit Distance
  - BERTScore
  - SentenceBERT
- To combine different techniques, we use the following techniques:
  - Reciprocal Rank Fusion (RRF)
  - Concatenation
- To evaluate the techniques, we use the $F_\beta$-score, $\beta = 0.25$.

# Results

## Named Entity Retrieval

- From regest, we took a stratified sample of 21 entities:
  - Shortest and longest (6)
  - German, Czech, and Latin (9)
  - People and places (6)
- We annotated top ten occurrences of every entity in books.

| | Precision | Recall | Fβ-score |
|---|---|---|---|
| Manatee | **100%** | 17% | **78%** |
| Fuzzy regexes | 79% | **24%** | 69% |

# Methods

## Named Entity Recognition

- As our baseline, we use Babelscape/wikineural-multilingual-ner.
- We train our recognition models by fine-tuning xlm-roberta-base.
- To train recognition models, we use three datasets:
  - Book texts (unsupervised, MLM)
  - Entities in regests (supervised, NER)
  - Entities in books (supervised, sparse, NER)
- To train recognition models, we also use two types of schedules:
  - Sequential (first MLM, then NER)
  - Parallel (both MLM and NER)
- To evaluate the techniques, we use the $F_\beta$-score, $\beta = 0.25$.

# Results

## Named Entity Recognition

- Results on entities in regests (unilingual)

| Precision (P) | | | Recall (R) | | | Fβ-score | | |
|---|---|---|---|---|---|---|---|---|
| LOC | PER | All | LOC | PER | All | LOC | PER | All |
| 58% | 88% | 76% | 54% | 40% | 43% | **58%** | **82%** | **75%** |
| | | | | | Baseline | 6% | 20% | 13% |

- Results on entities in books (sparse)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 31–100% | 29–100% | 30–100% | 71% | **66%** | **68%** | 32–100% | 30–100% | 31–100% |
| | | | Baseline | **77%** | 60% | 67% | | |



MUNI FI

# Future Work

- Produce a proper test dataset:
  - Use recognition models to automatically tag missing entities in books.
  - Have historians from ARTS to manually check the resulting dataset.
- Select our best model and tag missing entities in training data.
- Train larger model on completed training data.
- Refer about our results in a proceedings article:
  - NOVOTNÝ et al. The Retrieval and Recognition of Named Entities in Medieval Texts. In Horák et al. RASLAN 2022. Brno.


FACULTAS ARTIS INFORMATICÆ
UNIVERSITAS MASARYKIANA