P 39 - Extracting surnames from old church books

Braian Dias  
Tahira Iqbal  
Asim Seyidov  
Aitor Urkiza  
PO: Sulev Reisberg (University of Tartu)
Problem

- **Old church handwritten books** in Estonia are available on the internet.
  - Birth, marriage, death registries
- [https://www.ra.ee/dgs/addon/nimreg/about.php](https://www.ra.ee/dgs/addon/nimreg/about.php)
Problem

How to find useful information in these books?

- Some books have been indexed by names or surnames.
  - Manual work
Our task

- Apply **machine learning techniques** to extract names from the books to know on what page they appear
- **Book**: Personnel Book V (Kõnnu, Kiiu, Kolga, Anija, ...; EAA.1216.1.275; 1888-1902
- **No labeled data**
- **Objective:**
  - SURNAME 1, LINK TO FRAME 1, TITLE OF DOCUMENT 1
  - SURNAME 2, LINK TO FRAME 2, TITLE OF DOCUMENT 2
Our task

Handwritten Text Recognition (HTR)

Text Detection (Segmentation)

Handwritten Recognition

Surname detection

Kassispää
2) Hinrik Einman 9 III 64; 31 X 82
3) Loowisa 6170; 8 III 87

Kassispää
2) Hinrik Einman 9 III 64; 31 X 82
3) Loowisa 6170; 8 III 87
Challenges in HTR

- Tremendous variability and ambiguity of strokes;
- Inconsistency in writing style;
- Oftentimes text is not written in a strictly straight line
- **Cursive handwriting difficult to understand**
- Text in handwriting can have a variable rotation
- **Collecting a high-quality dataset with labels is relatively expensive**
- Labeled data not available

Proposed approaches (HTR)

Supervised HTR Model (CNN+RNN):

- Tesseract with LSTM (https://github.com/tesseract-ocr/tessdoc #40-with-lstm)
- Not enough data available to train/fine tune the model

Existing HTR models/tools:

- TRANSKRIBUS - Focus on historical documents (https://readcoop.eu/transkribus/)
Proposed approaches (HTR)

Few-shot learning

- 5 samples per symbol required
- Complex. Code related to the paper wasn’t available*
Proposed Solution
Our Solution

Hypothesis:

(1) Can we use existing HTR models for extracting text in church books?
(2) Can we improve these models by fine tuning them with task-specific data?
Our Solution

TRANSKRIBUS

Recognized Text

Woldg Adams
V. Johan
1.16. Annadeb. Lihaá
S. Elmar
S. Wöldemar

Filtered Text

Adams
Elmar

Estonian names
Our Solution

**TRANSKRIBUS**

**Recognized Text**

- Woldg Adams
- V.Johan
- 1.16. Annadeb. Lihaá
- S.Elmar
- S.Wöldemar

**Filtered Text using DB (similarity search)**

- Adams
- Elmar

~1000 surnames
Surnames changed with time...
- "w" -> "v"
Our Solution

TRANSKRIBUS

Recognized Text

Woldg Adams
V.Johan
1.16. Annadeb. Lihaá
S.Elmar
S.Wöldemar

Regex based Filter

Adams
Johan
Lihaá
Elmar
Wöldemar

Estonian names
Dataset

Manual annotation of surnames using Transkribus

- Focus on annotating full names where surname is present
- **30 pages** randomly selected from 334 pages (~9%)
  - 2555 Lines
  - 4928 words
- Leveraged existing models for text detection
Method (1) - Baseline

Selected a few HTR models available in Transkribus

Tested models (baseline):

- Transkribus German Kurrent
  - TRANSKRIIBUS GERMAN KURRENT XIX_M2
  - 19th Century scripts
- National Archives of Finland 19th century
  - NAF Court Records M10
  - Swedish 19th century
- Estonian Court Records 19th century
  - Estonian 19th century
Method (2) - "augmented" models

Improve existing models by using our manually annotated pages

Transkribus has a feature to train a new model on top of an existing one

Models (ours):

- **Estonian augmented v1**
  - Based on Estonian Court Records 19th century
  - + 20 annotated pages from church books

- **German augmented v1**
  - Based on TRANSKRIBUS GERMAN KURRENT XIX_M2
  - + 20 annotated pages from church books

- **Finish augmented v1**
  - Based on NAF Court Records M10
  - + 20 annotated pages from church books
Final pipeline

TRANSKRIBUS

Recognized Text

Woldg Adams
V.Johan
1.16. Annadeb. Lihaá
S.Elmar
S.Wöldemar

Regex based Filter (python script)

Adams
Johan
Lihaá
Elmar
Wöldemar

XML

Code available here:
Results

Test set: Pages from our annotations considering only surnames

- 87 surnames from 10 pages
- Accuracy: correct matches/surnames (GT)

<table>
<thead>
<tr>
<th>model</th>
<th>matches</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estonian (base)</td>
<td>5</td>
<td>5.75%</td>
</tr>
<tr>
<td>Estonian (ours)</td>
<td>41</td>
<td>47.13%</td>
</tr>
<tr>
<td>Finish (base)</td>
<td>21</td>
<td>24.14%</td>
</tr>
<tr>
<td>Finish (ours)</td>
<td>43</td>
<td>49.43%</td>
</tr>
<tr>
<td>German (base)</td>
<td>25</td>
<td>28.74%</td>
</tr>
<tr>
<td>German (ours)</td>
<td>4</td>
<td>*4.60%</td>
</tr>
</tbody>
</table>

* German model requires more epochs to train.
All models were trained with 30 epochs.
Conclusions

● Using **existing HTR models** for text recognition seems promising, however:
  ○ Fine tuning a model for church books requires additional data
    ■ Annotations requires domain knowledge about the scripts (letters and symbols from 1800)
    ■ Accurate annotations take a lot of time
● Our approach outputs a lot of **false positives** since post processing on the transcriptions is based on regular expressions
  ○ If HTR outputs accurate text is possible to use **NLP techniques** to extract surnames
    ■ Also might require text annotations
● **Few-shot learning** was considered, but it would require more time to implement
Future work

- NLP for improving surname detection on extracted text
  - POS tagging, Named Entity Recognition
- Few-shot learning
  - Research on low resource handwriting recognition
    - Only 5 samples per symbol required
Questions?