Comparing human and machine performances in transcribing 18th century handwritten Venetian script

Sofia Ares Oliveira

Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne

DH2018, June 2018, Mexico City
Objective:

Make collections of digitized archival records accessible through **textual search**

Usually two options for transcription of large collections:

- **Crowdsourcing**
- **Automatization**

Can automatic transcription be as good as crowdsourcing approach?
Dataset

Subset of 18th century fiscal documents from the Venetian State Archives

Catastici delle parrocchie
Subset of 18th century fiscal documents from the Venetian State Archives

Catastici delle parrocchie

Indici
Dataset

Subset of 18th century fiscal documents from the Venetian State Archives

Catastici delle parrocchie
Indici
Quaderni dei Trasporti
Dataset

Subset of 18th century fiscal documents from the Venetian State Archives

Catastici delle parrocchie
Indici
Quaderni dei Trasporti
Indici
23 000 units (image segments) manually transcribed by trained archivists

54 200 Venetian names of persons and places
Machine transcription

vs

Human transcription
Neural network architecture: CRNN


Height is fixed for all the image segments (but image ratio is kept)

Data augmentation (contrast, intensity, rotation)

20,712 image segments
48,628 words in total
8,848 vocabulary items
‘A-Za-z’ characters + a few symbols for punctuation
Evaluation

2 317 image segments
5 559 words in total
2 157 vocabulary items

Character Error Rate (CER) : 0.0804
Machine transcription vs Human transcription
Evaluation of the average performance of Italian-speaking transcribers

Platform: CrowdFlower (now Figure Eight)

Task: Transcribe text in image segment, taking into account capitals and punctuation

Data: 2,317 image segments from test set
Evaluation of the reliability of transcribers' answers during the experiment:
- 103 evaluation units
- 0.6 accuracy required

36 transcribers remained after selection

8,674 valid transcriptions to analyze
<table>
<thead>
<tr>
<th></th>
<th>CER system</th>
<th>CER amateur</th>
<th>WER system</th>
<th>WER amateur</th>
</tr>
</thead>
<tbody>
<tr>
<td>No formatting</td>
<td>0.0804</td>
<td>0.1328</td>
<td>0.2709</td>
<td>0.4318</td>
</tr>
<tr>
<td>With formatting*</td>
<td>0.0718</td>
<td>0.1047</td>
<td>0.2551</td>
<td>0.3507</td>
</tr>
</tbody>
</table>

*formatting = replace capital letters by lowercase and remove punctuation
On going work

Extend the recognition to **abbreviation** symbols

P: francesco Zonello quondam alessandro  
GT: francesco Tonello quondam alessandro

P: fratelli Fappa detta Vattin quondam Giacomo  
GT: fratelli Foppa detti Vatta quondam Giacomo

P: reverendo don Iacomo, ed antonio fratelli Picati quondam Leonardo  
GT: reverendo don Iacomo, ed antonio fratelli Pujati quondam Leonardo

P: 1794 primo maggio  
GT: 1794 primo maggio
The system has lower CER and WER than amateur transcribers’ average on 18th century Venetian script

→ Sufficiently reliable to use for searching purposes

→ New prospects for analyzing and study large collections of documents
Sofia Ares Oliveira
sofia.oliveira@epfl.ch

Frederic Kaplan
frederic.kaplan@epfl.ch

Digital Humanities Laboratory
dhlab.epfl.ch

Venice Time Machine
vtm.epfl.ch

github.com/solivr/tf-crnn