Handwritten Text Recognition (HTR) model for historical documents from 17th to 20th centuries – Using TrOCR

Marttila Riikka[1], Joska Sanna[1], Lipsanen Mikko[1], Föhr Atte[1], Jokipii Ilkka[1]
[1] National Archives of Finland
firstname.lastname@kansallisarkisto.fi

Handwritten historical documents remain an important part of research materials in different fields even today. When historical documents are digitized, modern text analysis methods can be applied to make them easier to use and analyse. However, texts must first be recognized from images containing text and converted into machine-typed form. In a research project conducted at the National Archives of Finland, we utilize the pretrained Transformer-based Optical Character Recognition (TrOCR) model developed by Microsoft (Li et al. 2022). It combines an image Transformer encoder and a text Transformer decoder for optical character recognition, replacing traditional CNN- and RNN-based approaches and eliminating the need for additional language models for post-processing accuracy. TrOCR is pre-trained on synthetic data and fine-tuned on human-labeled datasets, demonstrating superior performance on both printed and handwritten text recognition tasks.

In this research, we aim to compare the performance of various HTR models developed specifically for the handwriting styles of individual centuries against a super model trained on a comprehensive dataset from the 1600s to 1900s. Another goal of the research is to train an HTR model to perform with sufficient accuracy on documents in both Finnish and Swedish languages. With the help of high-performing HTR models, The National Archives can make handwritten historical documents more accessible and easier to use as source materials in many fields, such as historical or linguistic research.

This research has access to 26800 pages of annotated data. Annotation here refers to the transcription of texts and the marking lines around text lines. On average, one page consists of 30 lines of text. The data is randomly divided into training, validation, and test datasets. The training dataset is used for training the HTR model, while the validation set is automatically used by TrOCR for model validation to identify the best model configuration. The test dataset is used to compare different models against each other. The test data has been randomly selected and weighted across different centuries and languages (Swedish and Finnish) to ensure a sufficiently representative sample from each century and both languages. Different models are compared, and model accuracy is evaluated using the Character Error Rate (CER) value.

Keywords: handwritten text recognition, machine learning, historical documents

References

Li, M.; Lv, T.; Chen, J.; Cui, L.; Lu, Y.; Florencio, D.; Zhang, C.; Li, Z.; Wei, F. 2022. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *arXiv: 2109.10282v5*