Handwritten Text Recognition (HTR) model for historical documents from 17th to 20th centuries – Using TrOCR

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

Poster - Abstract

Handwritten historical documents remain an important part of research materials in different fields even today. When historical documents are digitized, modern text analysis methods can be applied to make them easier to use and analyse.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-readable form, and for a long time this has not been possible for handwritten texts. Without the advancements in (open source) machine learning methods and computational power in recent years, handwritten text recognition on a larger scale would not have been possible (Muehlberger et al. 2019). 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.

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. The goal is to train HTR models to perform with sufficient accuracy on documents in both Finnish and Swedish languages. As an important part of the strategy of the National Archives of Finland high-performing HTR model development can make handwritten historical documents more accessible and easier to use as source materials in many research fields (Lahtinen & Katajisto 2020, Paju et al. 2020).

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 and weighted across different centuries and languages (Swedish and Finnish) to ensure a sufficiently representative samples from each century and both languages. 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. Different models are compared, and model accuracy is evaluated using the Character Error Rate (CER) value.

Keywords: digital humanities, handwritten text recognition, historical research, historical documents, machine learning

REFERENCES

Lahtinen, A. & Katajisto, K. (2020). *Handwritten text recognition opens up a treasure trove of information on Finnish society. In J. Nuorteva & P. Happonen (Eds.),* The National Archives of Finland Strategy 2025: Perspectives for the Future (pp.18-19).

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

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., . . . Zagoris, K. (2019). *Transforming scholarship in the archives through hand-written text recognition: Transkribus as a case study.* Journal of Documentation, Vol. 75 No. 5, pp. 954-976. https://doi.org/10.1108/JD-07-2018-0114

Paju, P., Oiva, M., & Fridlund, M. (2020*)*. Digital and distant histories: Emergent approaches within the new digital history. In M. Fridlund, M. Oiva, & P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history* (pp. 3–18). Helsinki: Helsinki University Press. https://doi.org/10.33134/HUP-5-1

Acknowledgements

We would like to thank everyone who participated the project with special thanks to all the annotators as well as the volunteers and researchers who wrote the transcriptions.