Post-OCR Correction of Digitized Swedish Newspapers with ByT5

Viktoria Löfgren
Department of Computer Science and Engineering
Chalmers University of Technology
viktoria.lofgren@live.se

Dana Dannélls
Språkbanken Text
University of Gothenburg
dana.dannells@svenska.gu.se

Abstract

Many collections of digitized newspapers suffer from poor OCR quality, which impacts readability, information retrieval, and analysis of the material. Errors in OCR output can be reduced by applying machine translation models to 'translate' it into a corrected version. Although transformer models show promising results in post-OCR correction and related tasks in other languages, they have not yet been explored for correcting OCR errors in Swedish texts. This paper presents a post-OCR correction model for Swedish 19th and 20th century newspapers based on the pre-trained transformer model ByT5. Three versions of the model were trained on different mixes of training data. The best model, which achieved a 36% reduction in CER, is made freely available and will be integrated into the automatic processing pipeline of Språkbanken Text, a Swedish language technology infrastructure containing modern and historical written data.

1 Introduction

The OCR (Optical Character Recognition) quality of printed documents in general and historical documents in particular is often low. Historical documents often suffer from stains, faded print, and ink bleed-through from other pages, which leads to poor OCR accuracy. This, in turn, causes a range of challenges for Natural Language Processing (NLP) systems such as information retrieval and analysis. One way to achieve higher accuracy is to apply a post-OCR correction method to the OCR output. In this context, post-OCR can be compared to machine translation where the input is a set of character strings in one form that should be mapped onto the corrected form (Nguyen et al., 2021).

Since the transformer architecture was introduced in 2017 (Vaswani et al., 2017), it has pushed the state-of-the-art in many NLP tasks, including machine translation. It has also led to the emergence of large pre-trained models. One of these is ByT5, which is pre-trained on the large web-scraped multilingual dataset mC4 (Xue et al., 2021b).

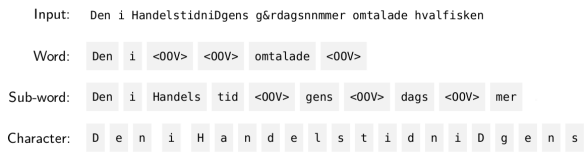

Figure 1: Word, sub-word and character-level tokenization of the sequence. *Den i HandelstidniDgens g&rdagsnnmmer omtalade hvalfisken*.

ByT5 operates on the character level, as illustrated in Figure 1. This approach preserves words that are not covered by the model's vocabulary due to, for example, OCR errors (*HandelstidniDgens*) or age (*hvalfisken*), at the cost of increased sequence length. Since the sizes of language models' vocabularies are fixed, a word-level model would map all out-of-vocabulary words (e.g., misspelled or obsolete words) to the same out-of-vocabulary token <OOV>, losing all information about these words. This character-level approach has reached state-of-the-art results in transliteration and grapheme-to-phoneme tasks (Xue et al., 2021a), diacritics restoration in 13 languages (Stankevičius et al., 2022), and post-OCR correction in Sanskrit (Maheshwari et al., 2022). However, it has not yet been explored for Swedish post-OCR correction.

The main contributions of this work are: (i) We demonstrate how effective a fine-tuned ByT5 is in the task of correcting OCR errors in 19th and 20th century Swedish newspapers. (ii) We show what effect does further training on data from books and other domains have on the model's performance on newspapers. (iii) We release a freely available post-OCR correction model with state-of-the-art performance on historical Swedish text, and a demo for testing the model. The model is available at https://huggingface.

Proceedings of LaTeCH-CLfL 2024, pages 237–242

co/viklofg/swedish-ocr-correction, with a demo at https://huggingface.co/spaces/viklofg/swedish-ocr-correction-demo.

2 Related Work

Previous work in post-OCR correction of Swedish historical text have explored both statistical and neural network based approaches. Persson (2019) used an SVM classifier in combination with a word list to detect and correct OCR errors in 17th to 19th century texts. Dannélls et al. (2021) proposed a method for increasing OCR accuracy of the Swedish newspapers by merging outputs from two OCR engines. Lundberg and Torstensson (2021) explored using the reference material prepared by Dannélls et al. to train an LSTM-based model from scratch, but found that it was not large enough to yield satisfactory results.

Another successful method for correcting OCR errors in Swedish historical text involves deep CNN–LSTM hybrid models (Drobac and Lindén, 2020; Brandt Skelbye and Dannélls, 2021). Drobac and Lindén (2020) utilized deep CNN–LSTM hybrid networks for the post-OCR task. They employed both Finnish and Swedish historical newspapers from 17th to 18th century, and reached state-of-the-art results for both. Brandt Skelbye and Dannélls (2021) experimented with mixed deep CNN–LSTM hybrid models directly on the character model within the OCR engine. While they have achieved state-of-the-art results for Swedish OCR, their method is not directly comparable to ours as it was not applied as a post-processing step.

In many other languages, post-OCR correction approaches based on neural machine translation have shown promising performance. Examples include the winner of the 2019 ICDAR competition (Rigaud et al., 2019), which was trained on ten European languages (but not Swedish). More recent examples include models for Finnish (Duong et al., 2021), Icelandic (Jasonarson et al., 2023), and English (Nguyen et al., 2020; Soper et al., 2021). Nguyen et al. (2021) note that while these models tend to outperform other techniques, they need a lot of training data to be successful. Nevertheless, the emergence of pre-trained transformer models, which require less training data and perform better than traditional methods such as LSTM networks, gives us hope that these models will overcome some of the previously reported limitations for Swedish OCR.

Dataset	Partition	Time period	Chars (k)	CER
Newspapers	Tesseract	1818–2018	6,957	4.86
	Abbyy Finereader		6,928	3.85
Literature		1836–2001	7,267	1.63
Blackletter	Swedish fraktur	1626–1816	282	17.61
	Then swänska Argus	1732–1734	259	19.06

Table 1: An overview of the datasets

3 Data

This project's main source of data is a manually transcribed subset of the National Library of Sweden's digitized newspapers.[1] In addition to this dataset, three other datasets are used: one dataset containing OCRed literature and two datasets containing OCRed blackletter texts.

All datasets come with a ground truth. An overview of all datasets is given in Table 1.[2]

Newspapers The newspaper dataset was prepared by Dannélls et al. (2021) and comprises almost 44,000 text segments identified by layout analysis of 400 Swedish newspaper pages printed between 1818 and 2018. The dataset includes two versions of each segment, one processed with Abbyy Finereader and one with Tesseract. Spanning two hundred years, the dataset is diverse in both typography and orthography. A majority of the 19th century pages are printed in blackletter typefaces, which often results in worse OCR accuracy and other kinds of errors compared to modern typefaces. The contemporary Swedish spelling was largely settled with the 1889 and 1906 spelling reforms (Pettersson, 2005), which means that the dataset contains both modern and historical spelling, for example *vad* and *hvad* ('what'), and *kvarn* and *qvarn* ('mill').

Literature The literature dataset consists of 79 titles of Swedish literature provided by the Swedish Literature Bank.[3] In total these texts amount to about 7.3M characters, making it slightly larger than the newspapers. It is contemporary with the newspaper dataset, but the OCR quality is generally much higher, likely because of higher print quality and simpler page layout.

Blackletter The blackletter dataset is a combination of two datasets prepared by Borin et al. (2016):

[1] https://tidningar.kb.se/

[2] A quantitative description of the size of the diachronical component of the dataset can be found in Brandt Skelbye and Dannélls (2021).

[3] https://litteraturbanken.se/

Swedish fraktur and *Then swänska Argus*. Both contain OCRed texts printed in blackletter typefaces along with the ground truth. *Swedish fraktur* contains texts from 199 pages from the collections of Gothenburg University Library. *Then swänska Argus* is a dataset consisting of 25 issues of the periodical of the same name by Olof von Dalin.

4 Methodology

4.1 Preparing the data

The main pre-processing step was to split the datasets into short samples. Careful consideration was taken to keep the OCR output and its ground truth aligned, since misaligned training samples may encourage the model to delete or insert text. Each pair of OCR output and ground truth was aligned line-by-line using a modified version of Myers' difference algorithm (Myers, 1986). In our version, we consider two lines 'equal' if their CER is low enough (the threshold was adjusted manually to suit each dataset), which compares the two texts and returns the lines that are present in both texts.

The aligned texts were split on line breaks into samples of typically 1-2 lines. These samples were filtered based on the following conditions: (a) both the OCR output and ground truth should be at least four characters long, (b) the CER should be below 50%, and (c) the ground truth may not contain @, which is used to indicate illegible text.

The newspaper samples were randomly assigned to three splits: train (70%), test (15%), and evaluation (15%). The two versions of each newspaper sample (one processed by Tesseract, one by Abbyy Finereader) were put in the same split to ensure that there was no contamination between the sets. The literature and blackletter samples were randomly assigned to two splits each: train (85%) and test (15%). These datasets were not used in evaluation, since the model's target domain is newspapers.

Dataset	Train	Test	Eval.
Newspapers	125,637	26,456	26,960
Literature	63,867	11,271	0
Blackletter	4,881	861	0

Table 2: Sizes of the three datasets (number of samples)

4.2 Fine-tuning setup

We fine-tuned three models using the following setup. The base model, `byt5-small`, was accessed through Huggingface's Transformers library (Wolf et al., 2020). The maximum input and output lengths were set to 128 UTF-8 bytes, which corresponds to slightly less than 128 characters of Swedish text.[4] The models were fine-tuned for three epochs using the Trainer API provided by Huggingface. Adafactor (Shazeer and Stern, 2018) was used as optimizer with a constant learning rate of 0.001, mimicking the setup used by Xue et al. (2021a) and Raffel et al. (2019) in fine-tuning ByT5 and T5, respectively. The batch size was set to 32, giving a total batch size of $128 \cdot 32 = 2^{12}$ tokens (bytes) per batch.

4.3 Models

Three models were fine-tuned: Model 1, Model 2 and Model 3. The only difference between them is what mix of the datasets, described in Section 3, they were fine-tuned on. The first model, Model 1, was trained on the newspaper dataset only, consisting of 126,000 training samples. Model 2 was fine-tuned on the newspaper and literature datasets, giving a total of 190,000 training samples. Since the literature dataset is contemporary with the newspapers, our ambition with this mix was to provide more examples of 19th and early 20th century Swedish. Model 3 was fine-tuned on the newspaper and blackletter datasets, giving a total of 131,000 training samples. Our ambition with this training mix was to provide more examples of typical errors in OCR of blackletter text, and in turn improve Model 1's performance on older newspapers.

4.4 Evaluation

Each fine-tuned model was evaluated on the 15% subset of the newspaper data. This evaluation set was aligned and filtered in the same way as the training data. The predicted corrections were computed using greedy decoding, i.e., at each decoding step, the highest-probability token was selected. The CER and WER were computed using the Python library `jiwer`.[5]

5 Results and Discussion

Table 3 shows the error rates of the evaluation set before and after processing with each model. All three models successfully reduced the error rates at

[4] ByT5 uses UTF-8 encoding, in which most characters occupy one byte, but non-ASCII characters such as *å*, *ä*, and *ö* occupy at least two bytes.

[5] Version 3.0.3., available at `https://pypi.org/project/jiwer/`, accessed November 6, 2023.

Period	CER (%)				WER (%)			
	BL	M1	M2	M3	BL	M1	M2	M3
1818–1859	8.39	4.39	4.63	**4.30**	32.46	**15.34**	15.80	15.54
1860–1899	4.04	**2.29**	2.61	2.38	16.51	**8.06**	8.63	8.22
1900–1939	2.60	2.01	1.97	**1.92**	11.24	7.03	7.08	**6.99**
1940–1979	1.46	1.39	1.49	**1.29**	6.45	4.43	4.54	**4.39**
1980–2018	0.83	**0.67**	0.75	0.73	3.74	**2.67**	**2.67**	**2.67**
1818–2018	3.20	2.06	2.20	**2.04**	13.21	**7.17**	7.42	7.23

Table 3: CER and WER of Model 1 (M1), Model 2 (M2) and Model 3 (M3) compared to baseline (BL).

both character and word level. This improvement can be seen in both modern and historical texts. Even though it can be assumed that ByT5 has not seen much historical Swedish in its pre-training data, the models did not seem to struggle disproportionately with correcting OCRed historical texts. It is possible that the error patterns are more predictable in the older material than the newer material, and thus easier to find and correct.

Over the entire set, the results in Table 3 show that Model 3 achieved the lowest CER of 2.04%, which corresponds to a 36% reduction from the baseline 3.20%. At the same time, Model 1 achieved the lowest WER, corresponding to a reduction of 46% (from 13.21% to 7.17%). These results are comparable to previously reported results in post-OCR correction of historical Icelandic texts using the larger `byt5-base` (Jasonarson et al., 2023), indicating that `byt5-small` may be sufficiently large for the task.

Although the differences between the three models' error rates listed in Table 3 are small, Model 1 and Model 3 tended to perform slightly better than Model 2. A possible explanation of this tendency is Model 2's relatively large portion of non-newspaper testing data. However, when inspecting the corrections produced by the models, we could not find any evident differences in quality. As an example, consider the evaluation sample shown in Figure 2. The three models agreed that *ko»unqen«* was incorrect, but only Model 3 managed to correct it to *Konungen* ('the king'). At the same time, Model 3 was unable to correct *alk* to *att* ('to').

The example in Figure 2 also displays the models' unwanted tendency to occasionally introduce new errors, for example *sä* → *få* ('get') instead of *sä* → *så* ('so'). In fact, the best model in terms of CER, Model 3, increased the CER in 7.7% of the evaluation samples. It is possible that the corrections could benefit from using another decoding strategy than greedy decoding.

OCR output (Model input)
— tz. M. ko»unqen« tillfriffFnanbe kor lock,
ligtwis nu sä fortgått alk H. M. den >6 för för,

Expected output
— H. M. konungens tillfrisknande har lyck-
ligtwis nu så fortgått att H. M. den 16 för för-

Model 1 output
— H. M. kommißens tillfrisknande kor lock,
ligtwis nu få fortgått att H. M. den 16 för för-

Model 2 output
— H. M. kommens tillfrisknande för lock-
ligtwis nu så fortgått att H. M. den 16 för för-

Model 3 output
— H. M. Konungens tillfrisknande kor lock-
ligtwis nu så fortgått all H. M. den 16 för för-

Figure 2: A sample from the evaluation set with corrections suggested by the three models. Errors and corrections are highlighted in bold.

6 Conclusion

In this paper, we present state-of-the-art results in post-OCR correction of Swedish 19th to 21th century newspapers. We fine-tuned three models from `byt5-small`, the smallest available version of Google's pre-trained character-level transformer model ByT5, using mixes of training data from different domains. The most successful model in terms of CER was fine-tuned on a mix of Swedish newspapers and blackletter texts. It achieved a 36% reduction in CER over the entire evaluation

set, but despite this, it was found to increase the CER in 7.7% of the evaluation samples. We made this model freely available and will integrate it into the automatic processing pipeline of Språkbanken Text,[6] a Swedish language technology infrastructure containing modern and historical written data. Further work aims to minimize the amount of introduced errors taking context information into account. We also intend to conduct several evaluations to learn more about the type of errors the model makes, in particular, on new digitized resources.

Limitations

The proposed model operates directly on text output from OCR engines. This makes it engine-agnostic and not reliant on any specific OCR output format, but may also limit its performance since it is unable to consider metadata such as character confidence scores. Without this knowledge, the model is equally prone to change a (possibly correctly recognized) character regardless of how confident the OCR software is.

Acknowledgements

This research was funded jointly by Nationella språkbanken and HUMINFRA, both received financial support by the Swedish Research Council (grant no. 2017-00626; and 2021-00176).

The authors would like to thank the Swedish Literature Bank for providing us with the literature dataset and its transcription. We would also like to thank the reviewers for their valuable comments on an earlier version of this paper.

References

Lars Borin, Gerlof Bouma, and Dana Dannélls. 2016. A free cloud service for OCR / En fri molntjänst för OCR. Technical report, University of Gothenburg, Gothenburg.

Molly Brandt Skelbye and Dana Dannélls. 2021. OCR processing of Swedish historical newspapers using deep hybrid CNN–LSTM networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online. INCOMA Ltd.

Dana Dannélls, Lars Björk, Ove Dirdal, and Torsten Johansson. 2021. A two-OCR engine method for digitized Swedish newspapers. In *CLARIN Annual Conference*. Linköping Electronic Conference Proceedings.

Senka Drobac and Krister Lindén. 2020. Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–17.

Quan Duong, Mika Hämäläinen, and Simon Hengchen. 2021. An unsupervised method for OCR post-correction and spelling normalisation for Finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 240–248, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Atli Jasonarson, Steinþór Steingrímsson, Einar Freyr Sigurðsson, Árni Davíð Magnússon, and Finnur Ágúst Ingimundarson. 2023. Generating errors: OCR post-processing for Icelandic. In *The 24rd Nordic Conference on Computational Linguistics*, pages 286–291, Tórshavn, Faroe Islands. University of Tartu Library.

Arvid Lundberg and Mattias Torstensson. 2021. *Deep learning for post-OCR error correction on Swedish texts*. Master's thesis, Chalmers University of Technology, Gothenburg.

Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. A benchmark and dataset for post-OCR text correction in Sanskrit. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6258–6265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eugene W. Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica*, 1:251–266.

Thi Nguyen, Adam Jatowt, Mickaël Coustaty, and Antoine Doucet. 2021. Survey of post-OCR processing approaches. *ACM Computing Surveys*, 54:1–37.

Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. 2020. Neural machine translation with BERT for post-OCR error detection and correction. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 333–336, New York, NY, USA. Association for Computing Machinery.

Simon Persson. 2019. *OCR post-processing of historical Swedish text using machine learning techniques*. Master's thesis, Chalmers University of Technology, Gothenburg.

Gertrud Pettersson. 2005. *Svenska språket under sjuhundra år*. Studentlitteratur.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

[6] https://spraakbanken.gu.se/en

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2019. ICDAR 2019 competition on post-OCR text correction. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1588–1593. IEEE Xplore.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv e-prints*.

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021. BART for post-correction of OCR newspaper text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online. Association for Computational Linguistics.

Lukas Stankevičius, Mantas Lukoševičius, Jurgita Kapočiūtė-Dzikienė, Monika Briedienė, and Tomas Krilavičius. 2022. Correcting diacritics and typos with a ByT5 transformer model. *Applied Sciences*, 12(5).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv e-prints*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.