Evaluating Transformers for OCR Post-Correction in Early Modern Dutch Comedies and Farces

Anonymous ACL submission

Abstract

In this paper, we investigate the performance on OCR post-correction in early modern Dutch of two types of transformers: large generative models and sequence-to-sequence models. To this end, we create a parallel corpus by automatically aligning OCR sentences to their ground truth from the EmDComF early modern Dutch comedies and farces corpus, and propose an alignment methodology that creates new segments based on combinations of newline splits. This improves the alignment between gold and OCR texts, which is essential for the creation of a high-quality parallel corpus. After filtering out misalignments, we fine-tune and evaluate both generative and sequence-to-sequence models. We find that mBART outperforms generative models for the automatic post-correction of early modern Dutch in the EmDComF corpus, correcting more OCR sequences and avoiding overgeneration.

1 Motivation & Related Work

Inspired by the success of generative large language models in a variety of NLP tasks, their usefulness has recently also been explored to automatically correct the output of Optical Character Recognition (OCR) models for historical documents. Evaluating 14 foundation models in zero and few-shot settings against 8 OCR post-correction benchmarks for manuscripts, newspapers, literary commentaries and other historical documents in different languages, time periods and transcription quality, Boros et al. (2024) found that generative models did not improve faulty transcriptions in their applied experimental settings. Moreover, they often degraded the transcription quality of texts. Conversely, Thomas et al. (2024) compared generative models for OCR post-correction after supervised fine-tuning (SFT) to prevalent sequence-to-sequence models for OCR post-correction on BLN600 (Booth et al., 2024), a dataset of 19th century British newspaper articles, and reported a Character Error Rate (CER) reduction of 54.51% after instruction-tuning generative models for a prompt-based approach to OCR post-correction.

Both Boros et al. (2024) and Thomas et al. (2024) recommend fine-tuning transformers on period-, genre- and quality-specific datasets to optimise OCR post-correction. Therefore, the models and results of their experiments do not directly transfer to datasets from other periods and languages. In the latter study, fine-tuned generative models were evaluated on 19th century English newspapers, which is relatively close to the mostly English web-crawled training data of generative models. In this paper, we compare fine-tuned generative models to sequence-to-sequence models for OCR post-correction in the early modern Dutch (1650-1725) OCRed texts of EmDComF (Debaene et al., 2024), which are less represented in the pre-training data of generative models. Moreover, this type of historical Dutch is characterised by orthographical variation and significant lexical and semantic shifts compared to modern standard Dutch, further complicating automatic text processing.

We first discuss the methodology for converting OCR and manually digitised texts into a parallel corpus through automatic sentence splitting and alignment in Section 2. After improving the default sentence alignment, we prepare the EmDComF dataset for automatic OCR post-correction, for which we train and evaluate SOTA systems in Section 3. Finally, we draw conclusions from our experiments on the particularities concerning correcting OCR output in EmDComF in Section 4.

2 Alignment

We make use of the EmDComF subset that has both OCR data, transcribed with Transkribus Print M1, and gold standard texts. This leaves us with 92 texts that originate from Census Nederlands

Toneel (CENETON) and 34 texts from Digitale Bibliotheek voor de Nederlandse Letteren (DBNL). As EmDComF consists of unstructured .txt files, we split the full-play texts of the OCR output into sentences using nltk (Bird et al., 2009) to create workable units. To achieve comparable gold units, this approach is also applied to the manually digitised versions of these texts. However, OCR and gold texts almost never have exact matching sentence lists, either due to typical OCR mistakes omitting punctuation or poor scanning quality. Another source of misalignments are occasional human errors in the gold texts. Due to this, we ignore casing and punctuation for the evaluation.

In this section, we present a dynamic alignment approach that improves the creation of an OCR parallel corpus. The goal of the presented alignment experiments is to make the parallel texts of EmDComF operational for automatic OCR post-correction. Hereto, we separate alignment mismatching from real OCR mistakes by introducing alignment after sentence-chunk splitting.

2.1 Methodology

In our experiments, we explore two different approaches for alignment. The baseline approach aims to match the full sentences as they were extracted with nltk (Bird et al., 2009). In addition, we propose an approach that splits the OCR sentences into smaller chunks based on newline characters so they can be better aligned with the gold text. After splitting a sentence into newline chunks, we create a variety of potentially alignable OCR segments by combining newline chunks that directly follow each other into reconstructed OCR sentences. This results in a significant increase in potential alignment matches for each gold sentence per text.

After creating these lists of gold sentences and alignment candidates for OCR sentences, we create sentence embeddings for each gold and all alignment candidates using all-mpnet-base-v2 (Reimers and Gurevych, 2019), a general-purpose sentence transformer. Then, we perform a semantic search based on cosine similarity to find the most similar OCR candidate for each gold fragment. OCR alignment candidates are limited to the same source text, i.e. to gold sentences of the same play, to avoid cross-text mappings. The same methodology for embedding and semantic search is applied to both alignment approaches.

2.2 Results

EmDComF contains texts from two databases that employ different formatting standards in their digitisation processes. Therefore, we provide results for both source databases. We remove exact duplicates of gold-OCR sentence pairs to minimise the impact of formatting standards, such as repeated character names or structure indications like acts and scenes. This results in a reduced combined set of 83,718 pairs, of which 60,874 originate from CENETON and 22,844 from DBNL, which allows us to focus the evaluation on the more relevant samples, i.e. the core text of the plays.

To evaluate the alignment, we calculate the cosine distance (based on the same sentence embeddings), character error rate (CER) and word error rate (WER), both edit distances relative to the length of the gold sentence, as well as normalised character error rate (nCER), relative to the longest OCR or gold sentence. The results in Table 1 show that our sentence-chunk splitting approach scores better on all metrics compared to the default sentence splitting approach. All score differences are statistically significant based on the p-values. This is supported by the box plots in Figure 1, in which the error rate distributions of our proposed approach are more compact in the combined corpora and have significantly fewer outliers.

Finally, after manual inspection we introduce a cosine distance threshold of 20% and normalised character error rate threshold of 40%, assuming that matches with larger distances must be mistakes in alignment. Our baseline sentence splitting approach indicates that 9147 pairs (CENETON: 6838, DBNL: 2309) were wrongly aligned. In comparison, our sentence-chunk splitting approach reduces the number of wrong alignments to 5632 (CENETON: 4134, DBNL: 1498). In addition to the mistakes in alignment, we also investigate the numbers of exact matches achieved by both approaches, where the alignment is flawless. Compared to the baseline approach, with 46,507 exact matches (55.55%), the reconstructed OCR sentences attain 50,208 exact matches (59.97%).

3 OCR post-correction

Based on our alignment approach, we combine both sets of OCR and gold standard sentences into a parallel corpus to conduct the OCR post-correction experiments in EmDComF. To guard the quality of these experiments, we remove cross-text duplicates

	CENETON				DBNL			
	cos_dist	nCER	CER	WER	cos_dist	nCER	CER	WER
sent_split	4.38	9.51	15.46	20.13	3.98	8.79	13.28	17.30
improved	2.91	6.66	7.87	11.78	2.79	6.37	7.27	10.67
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 1: Cosine distance, normalised CER, CER and WER before and after improved alignment in percentages. P-values indicate the statistical significance of the score difference per metric. See Appendix A for an example.

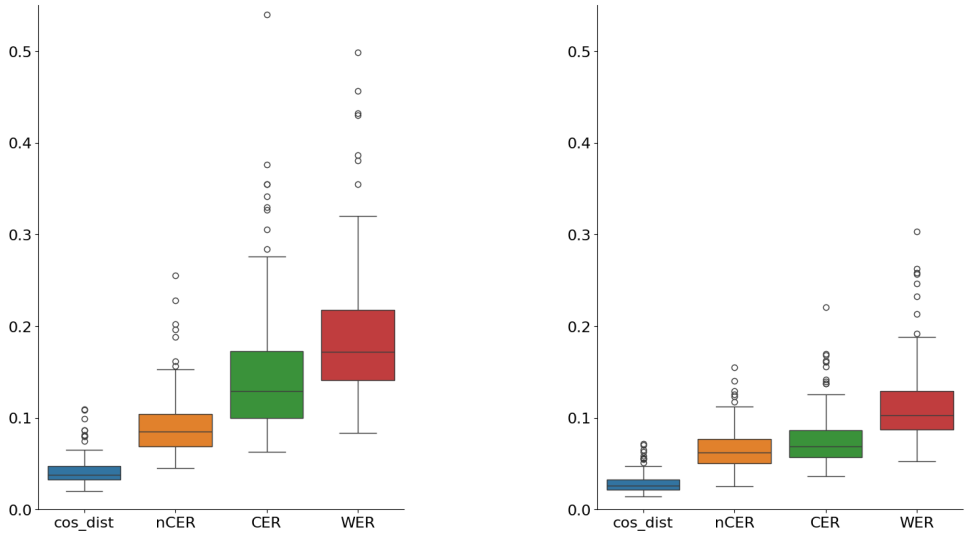

Figure 1: Comparison of sentence (left) and sentence-chunk (right) alignment on CENETON and DBNL combined.

and poor matches between sentence-chunk split OCR and gold (cfr. supra), since we cannot correct nor evaluate OCR sentences when they are aligned with the wrong gold standard. Doing this, we exclude wrong alignments from the experiments to correct OCR mistakes. The resulting OCR post-correction dataset consists of a train (52,894), test (14,693) and development (5,878) set, stratified for the percentage of exact string matches at 62.15%.

3.1 Methodology

For this corpus, we explore two distinct methodologies using transformer models. The first method involves fine-tuning a selection of pre-trained transformer models for sequence-to-sequence (seq2seq) modelling, which is considered the current state-of-the-art approach for OCR post-correction. The second method encompasses fine-tuning of large generative models, a novel approach that is gaining ground as the new state-of-the-art for many NLP problems. For seq2seq fine-tuning, we select mBART (Tang et al., 2020) as a strong multilingual model that has shown to work well for English.

To investigate the large generative models, we employ parameter-efficient fine-tuning, with QLoRa (Dettmers et al., 2023), which creates a low-rank decomposition of the weight matrix of the large model that can be trained for parameter-efficient fine-tuning. We make use of the supervised fine-tuning with trl (von Werra et al., 2020) based on a prompt that combines the OCR text with the gold output (Appendix B). As foundation models, we use the following selection of instruction-tuned models. We start from Llama 3 as a strong multilingual model and compare it to GEITje, a Dutch-specific model of a similar size and Fietje, a Dutch-specific model with a significantly lower number of parameters. Since these generative models are known to provide additional examples and explanations, we employ a set of post-processing rules to limit the model outputs to the correction of the original OCR sentence. These post-processing rules include the removal of the input of the prompt template, instruction-tuning tokens, and replacing outputs that are longer than the input text by 3 or more tokens with the baseline OCR text. For fair evaluation, the same rules are applied to the output of the seq2seq models.

3.2 Results & Discussion

The results from our experiments in Table 2 reveal that the mBART seq2seq model outperforms all generative models. This score difference is best reflected in mBART's lowered WER, meaning that

it succeeded best in concatenating, splitting, adding or removing OCR sequences, reducing error rates and increasing exact matches (Appendix C). This is further supported by the results on the subset focusing exclusively on samples requiring OCR post-correction (imperfect matches). Moreover, when considering the exact matches prior to OCR correction, we find that generative models produce exceedingly long sequences. To estimate overgeneration, we calculate the length difference between the gold sentences and the output after OCR post-correction. Whilst mBART has a mean difference of 0 on this subset of exact matches, Llama, GEITje and Fietje increase the character length by 0.05, 0.27 and 0.32 respectively. We hypothesise that generative models are more prone to overgeneration because they sequentially add tokens to the output and can therefore diverge more easily from OCR input texts than seq2seq models. We conclude that conceptually seq2seq models are more appropriate for this task as they focus on the input text when generating an output, which proves to be a significant advantage for processing both perfect OCR (exact matches with gold) and OCR with mistakes.

	nCER	CER	WER	#match
	testset (n=14,693)			
baseline	2.94	3.03	6.93	9131
Fietje	4.67	5.26	10.12	7575
GEITje	4.28	4.80	8.87	8204
Llama 3	2.85	2.96	5.94	9773
mBART	**2.81**	**2.90**	**5.76**	**9920**
	imperfect matches (n=5,562)			
baseline	7.77	8.01	18.31	0
Fietje	9.22	9.66	19.72	277
GEITje	8.56	8.99	17.15	692
Llama 3	7.26	7.52	14.83	906
mBART	**7.15**	**7.37**	**14.24**	**1053**

Table 2: Mean error rates and exact matches on the test set and the subset of imperfect matches. Baseline scores are calculated after alignment prior to post-correction.

Although Llama 3 is not directly pre-trained for Dutch, unlike GEITje and Fietje, it is the best performing generative model in our experiments, coming in as the second best performing model after mBART. Manual evaluation reveals that Llama 3 deviates less from the prompt template, whereas GEITje and Fietje require more system-specific rules to provide clean output. Additionally, GEITje and Fietje introduce noise into the OCR sentences more consistently than Llama 3, resulting in higher error rates and fewer exact matches than the baseline (raw OCR output). These results align with the work of (Boros et al., 2024), who investigated the performance of generative models in zero and few-shot settings. We thus find that fine-tuned generative models do not outperform seq2seq models for early modern Dutch OCR post-correction in the EmDComF corpus, which differs from the findings of Thomas et al. (2024). We hypothesise this is because their corpus is more closely related to the pre-training data of the generative models, facilitating the transfer to OCR post-correction for their type of historical data.

Finally, we recognise that Transkribus Print M1 performs well, even without post-correction, and establishes a strong baseline. As a result, there appears to be limited room for improvement through post-correction, making the task particularly challenging. Still, our best generative and seq2seq model significantly increase the exact matches after post-correction by 642 (4.36%) and 789 (5.37%), respectively, on the test set.

4 Conclusion

With this research, we advance the processing of automatically digitised historical texts by exploring the impact of sentence alignment and OCR post-correction for early modern Dutch in the EmDComF corpus. First, we propose an approach that combines OCR newline splits to create more valid candidates for sentence alignment when establishing an extensive parallel corpus. Though this approach is quite straightforward, it increases the exact matches from 56% to 60% and reduces wrong alignments from 10.93% to 6.73% on the corpus. Second, we evaluate the performance of SOTA models for OCR post-correction. The results of our experiments suggest that the mBART seq2seq model is the best performing approach for correcting OCR output, with lower error rates and an increase of 5.35% exact matches. Despite the solid performance of large generative models for OCR post-correction in late modern English newspapers in related work, sequence-to-sequence models remain the SOTA for early modern Dutch in EmDComF.

In conclusion, we consider the insights and methods from this work to be applicable to other projects focused on the creation and processing of corpora for digital humanities. In fact, improved digitisation could benefit many texts available online in scanned format, especially historical and linguistically non-normative datasets such as EmDComF.

Limitations

Firstly, the proposed methodologies were only tested on a single dataset. In our experiments we removed wrong alignments based on manually defined thresholds to create a parallel corpus. However, OCR mistakes may be the cause of wrong alignment and are not included in this study, though these examples may be relevant for post-correction. As a result, our scores might be a positive estimate due to these controlled conditions. Still, this group contains only 5,632 of the 83,718 pairs, leaving enough samples for analysis. In future work we will experiment with held-out train and test sets from different source databases, employing more out of distribution settings. Furthermore, the models can expectedly only be directly transferred to other historical Dutch corpora, as the performance on distantly or unrelated languages will likely be significantly different. We made use of sentence-level splitting of the texts in our methodology. Conversely, it would also be possible to group the texts into dialogues or paragraphs to allow language processing in larger and more meaningful contexts, which is possible for structured drama corpora like those available on DraCor. The automatic insertion of this type of structural knowledge into unstructured text is left for future work. Finally, we expect that generative models would likely work better on larger sequences than on the sentence level.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Callum William Booth, Alan Thomas, and Robert Gaizauskas. 2024. BLN600: A parallel corpus of machine/human transcribed nineteenth century newspaper texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2440–2446, Torino, Italia. ELRA and ICCL.

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

Florian Debaene, Kornee van der Haven, and Veronique Hoste. 2024. Early Modern Dutch comedies and farces in the spotlight: Introducing EmDComF and its emotion framework. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 144–155, Torino, Italia. ELRA and ICCL.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Appendix

A Alignment

{gold} **en wyl ik dat rapsody ken**

{sent_split} gitized by google \n 27 \n 28 \n beslikte swaantje \n **en wyl ik dat rapsody ken**

{improved} **en wyl ik dat rapsody ken**

{translation} and while I know that rapsody

B Prompt Template

{user} Correct the OCR errors in the provided text.
Not all texts contain errors.
Text: {INPUT_OCR}
{ass}### Corrected Text: {CORRECTED_OCR}

C Post-Correction

{gold} **zou** uw juffer weg weezen

{base} **ou** uw juffer weg weezen

{mBART} **zou** uw juffer weg weezen

{Llama 3} **ou** uw juffer weg weezen

{GEITje} **nou** uw juffer weg weezen

{Fietje} uw juffer weg weezen

{translation} shall your missus leave