Unsupervised Multi-View Post-OCR Error Correction With Language Models

Harsh Gupta† **Luciano Del Corro**† **Samuel Broscheit**‡* **Johannes Hoffart**† **Eliot Brenner**†
†Goldman Sachs
{Harsh.Gupta, Luciano.DelCorro, Johannes.Hoffart, Eliot.Brenner} @ gs.com
‡University of Mannheim
broscheit@informatik.uni-mannheim.de

Abstract

We investigate post-OCR correction in a setting where we have access to different OCR views of the same document. The goal of this study is to understand if a pretrained language model (LM) can be used in an unsupervised way to reconcile the different OCR views such that their combination contains fewer errors than each individual view. This approach is motivated by scenarios in which unconstrained text generation for error correction is too risky. We evaluated different pretrained LMs on two datasets and found significant gains in realistic scenarios with up to 15% WER improvement over the best OCR view. We also show the importance of domain adaptation for post-OCR correction on out-of-domain documents.

1 Introduction

Digital scans of printed paper are still one of the main sources of digitized text across industries, libraries and governmental organizations. Scanned documents need to be processed by Optical Character Recognition (OCR) systems in order to be consumed by natural language processing (NLP) pipelines. Unfortunately, OCR errors are pervasive and input noise can severely hinder downstream NLP applications, e.g., search (Stein et al., 2012), neural machine translation (Belinkov and Bisk, 2018) or NLU in general (Kumar et al., 2020).

Prior work used text generation techniques or redundancy in *similar* passages for OCR error correction, which is not appropriate in cases of low corpus redundancy or weak document contextual information. For example, this may pose a risk in sensitive documents in the legal or financial domain, where documents tend to be templatized and specific information, such as legal entities or numbers (e.g., interest rates or amounts), are specific to single documents and cannot be safely inferred. Other prior work uses multiple OCR views of a document and reconciles them in a supervised way, requiring expensive and difficult to acquire training data.

OCR_1:	Total	0 ft	suppl ies :	23 .64
OCR_2:	Total	0 %	suppl tea	.3 .64
Reconciled:	Total	0 %	suppl ies :	23 .64

Figure 1: Example ICDAR dataset. Output from two popular OCR systems (OCR_1 and OCR_2) and the reconciled version generated by our approach.

Therefore, we investigated post-OCR correction with multiple OCR views on the same document in an unsupervised way. A key assumption of this work is that different OCR views make mistakes in different parts of the input document. To create a better OCR view from multiple OCR inputs we use a language model (LM) to pick the most probable reconciliation. See Figure 1 for an example. The key question of this study is if it is possible to use a LM to reconcile the OCR views such that their combination contains fewer errors than the individual views. The advantage of the proposed approach is that (i) it does not require supervision, and that (ii) it is well suited for risky scenarios.

We explore two different settings: (i) using an off-the-shelf pre-trained LM (i.e., GPT/2 family and a n-gram model), and (ii) domain adaptation of the LM. We evaluated our approach on two datasets that we adapted for our experiments[1]. The RETAS dataset (Yalniz and Manmatha, 2011), consisting of 20 English books with a total of 100 OCR views, and the more challenging ICDAR Scanned Receipts dataset (Huang et al., 2019), with 625 scanned receipts. Our results indicate that the proposed unsupervised approach is able to improve over the individual OCR systems. On RETAS we measured a 8% gain over the best OCR view and on ICDAR a gain of 15% over the best OCR view. We also found that domain adaptation is crucial for

* work done during an internship at Goldman Sachs

[1] https://github.com/HarshGupta11/ocr_correction_refiner

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8647–8652
November 7–11, 2021.

documents from domains distinct from the LM's training data, as we can show that domain adaptation improved the error rates on critical numerical data.

2 Related Work

Multi-Input post-OCR correction. Using ensemble methods and voting schemas across multiple inputs are well established strategies for post-OCR error processing (Lopresti and Zhou, 1997; Yamazoe et al., 2011; Lund et al., 2013; Xu and Smith, 2017; Dong and Smith). The inputs may come from different views of the same document (Lopresti and Zhou, 1997; Lund et al., 2013) or from corpus redundancy using *similar* passages on a large corpora (Dong and Smith; Xu and Smith, 2017). Here we follow the first approach because corpus redundancy may be risky in some settings. For instance, in low variance documents, such as financial documents) the important information is document specific (e.g., amounts, interest rates, entities, etc.); extracting those values from other *similar* passages would lead to risky errors in the crucial document bits.

Input reconciliation. Earlier single-system multi-input approaches were relying on multiple scans of the same documents (Lopresti and Zhou, 1997), or alterations to the original image to force (Lund et al., 2013) alternative OCR outcomes. The reconciliation was performed either by voting schemas or direct supervision. Even those systems relying on document redundancy used supervision (Schulz and Kuhn, 2017; Dong and Smith) with manual or automatically generated training data. Unlike previous methods we are the first to use large LMs to reconcile the inputs in an unsupervised way. Our work is in this aspect similar to (Xu and Smith, 2017) as they used a character-based ngram LM in addition to the majority voting to decide among the similar passages, however they also rely on *similar* passages to generate redundancy.

Explicit correction. Text generation is a standard technique in post-OCR correction (Xu and Smith, 2017; Schulz and Kuhn, 2017; Amrhein and Clematide, 2018; Richter et al., 2018; Dong and Smith; Lyu et al., 2021). Neural approaches, for instance (Amrhein and Clematide, 2018; Dong and Smith; Nguyen et al., 2020; Lyu et al., 2021) use an encoder decoder architecture which takes the OCR'ed text as input and generates the corrected version. In our scenario these strategies can lead to severe problems. For instance, in the case of numerical values, contextualized LMs will probably be able to generate a numerical value, but this value will be arbitrary in the context of the specific document (e.g., 3,50%, 5%, etc.). An unreadable number is in this case better than a wrongly readable one as the error will be more easily visible for both humans and information extraction systems.

3 Background

3.1 Post-OCR error correction

Post-OCR error correction is the task of correcting the errors generated during the OCR process. It involves two main challenges: (i) the detection of the errors in text, and (ii) the correction of those errors. In general, the first challenge is non-trivial. For example, if there is only a single OCR input to the correction module, OCR errors might not be obvious corruptions (e.g., OCR_2 in Figure 1) which can result in text at least superficially readable, so even the location of the error is not obvious.

By using multiple OCR inputs of the same document this challenge can be sidestepped. Thus, instead of solving error detection we only have to address the much more specialized problem of determining the differences between multiple versions or "views". Spotting the differences is an established problem, with multiple techniques available, such as Yalniz and Manmatha (2011).

3.2 Language Models

In general, a *language model* (LM) may refer to any parameterized method of assigning a probability to a sequence $t_1, \ldots, t_k$ of tokens:

$$p(t_1, t_2, \ldots, t_k) = \prod_{n=1}^{k} p(t_n | t_1 \ldots t_{n-1})$$

In what follows we use a normalized form of the probability known as *perplexity*, defined as the inverse of the k-th root of the probability of the sequence. Note that we can distinguish two types of LMs, n-gram and neural, which are distinguished by whether they estimate the probability factor $p(t_k | t_1 \ldots t_{k-1})$ by simple statistical methods or by a neural network. For the analysis in this work we use 3 models of the GPT (Radford and Narasimhan, 2018; Radford et al., 2019) family, plus a 3-gram model trained on Wikipedia.

4 Proposed Approach

Our approach focuses on a setting with multiple OCR views where the challenge is to pick the best segments of each view where the aligned OCR inputs differ. Given two OCR views of the same document our approach consists of the following steps: (i) align the OCR outputs and spot differences, and then (ii) score the different choices of those differences to pick the best solution.

Step 1: Spot differences. Suppose we have two sequences $S_1 = c_{1,1}, \ldots, c_{1,n}$ and $S_2 = c_{2,1}, \ldots, c_{2,n}$, where $c_{i,j}$ is the j-th character of sequence i. Further we denote $d_{i,j}$ as the j-th chunk in sequence i which is not present in the other sequence, while e_i is a chunk shared by both inputs.

Finding the longest common subsequence between S_1 and S_2 is equivalent to finding the shortest edit script (Myers, 1986). Denote the shortest edit script by $\text{Diff}(S_1, S_2)$. For example, suppose that it takes the form $\text{Diff}(S_1, S_2) = e_1[d_{1,1}, d_{2,1}]e_2[d_{1,2}, d_{2,2}]e_3$. $[d_{1,j}, d_{2,j}]$ indicates that to transform S_1 into S_2, one must delete $d_{1,j}$ and insert $d_{2,j}$. In the example the number of differing chunks $\text{length}(\text{Diff}(S_1, S_2)) = 2$.

Each difference yields a binary choice to either pick $d_{1,j}$ or $d_{2,j}$. All possible outputs that can be produced from the pair of inputs (S_1 and S_2) is the set of all root-leaf paths in a binary tree.

Step 2: Score interpretations. We score all of the possible root-leaf paths with an LM's perplexity and take the *argmin*. If followed naively, this leads to an exponential computational complexity in the height of the tree. Therefore we adopt beam search with beam width β (Russell and Norvig, 2002).

5 Experiments

The goal of the experiments is to understand if the proposed approach is able to generate a combined OCR view which contains fewer errors than the individual inputs.

5.1 Datasets

We have the following requirements: (i) We need multiple OCR views of the same document and (ii) a ground truth for evaluation. As there were no public datasets that meet those requirements we adapted the RETAS dataset (Yalniz and Manmatha, 2011) and the more challenging ICDAR 2019 Competition on Scanned Receipt OCR (Huang et al., 2019). Other more standard datasets for post-OCR correction (Chiron et al., 2017; Rigaud et al., 2019) were designed for text correction approaches, and were not suitable for our setup.

RETAS Originally created for text alignment. We used 16 English books with 96 OCR views in total. The ground truth for each book comes from the Gutenberg Project and since the OCR views are generated from different editions of the book there are alignment miss matches that we solved in the following way: (i) we divided the ground truth in chunks of 200 characters, (ii) we discarded those chunks for which the aligned OCR views have a length discrepancy of more than 10%, (iii) we removed chunks across views with ground truths discrepancies, and (iv) we removed chunks that were not present across views. In total we discarded around 40% of the data. We made a 60:20:20 split for domain adaptation, validation and test respectively. The datasets were split per book to avoid leakage into the test set. We use 9 books for domain adaptation, 3 for validation and 4 for testing.

ICDAR 2019 Scanned Receipts. It contains 625 receipt images. Each image is annotated with text bounding boxes and the transcript of each text. We extracted a total of 18,228 lines, out of which we have perfect alignments between ground truth and the two OCR systems for 15,905. The rest were discarded. To generate the different OCR views, we processed the images with two popular OCR systems and discarded those images not processed by either of the systems, resulting in 533 receipts. We used 319 for domain adaptation, 107 for validation, and 107 for testing, randomly sampled in 5-fold cross validation.

5.2 Experimental Setup

Models. We use three autoregressive models: GPT, GPT2 and GPT2XL, and trained a 3-gram model on Wikipedia as baseline. One tunable hyperparameter is the maximum number of tokens in the prefix and suffix around the difference. We tuned this on the validation sets and used 50 characters on each side of the OCR difference for both datasets. For the beam search we use beam search with $\beta = 6$.

Domain adaptation. A typical use of neural LMs is to fine-tune them in a NLP task. This often involves two stages: unsupervised pre-training for domain adaptation and supervised task fine-tuning. For OCR correction, we do not want to assume there is any supervision, because that would require a corpus of OCR'ed text and manually transcribed text in the domain of interest. However, it can

Dataset	b	w	GPT2	GPT2-DA	GPT2XL
RETAS	3.74	7.36	**3.43**	3.50	3.45
ICDAR 2019	46.82	102.22	45.99	**40.81**	45.08

Table 1: WER of each OCR view per Dataset, Best (b) and Worst (w), best WER LMs, with and without domain adaptation (DA).

Model	Best+2nd		Best+Worst		Random	
	b	w	b	w	b	w
without domain adaption						
3-gram	-1%	49%	-13%	43%	-32%	33%
GPT	5%	52%	-6%	46%	-14%	42%
GPT2	8%	53%	-4%	47%	-10%	44%
GPT2XL	8%	53%	-4%	47%	-10%	44%
domain adaptation						
GPT	3%	51%	-8%	45%	-21%	39%
GPT2	6%	52%	-2%	48%	-11%	44%

Table 2: RETAS. WER relative improvement over best OCR view (b) and worst view (w). Combining Best+2nd, Best+Worst and iteratively combining all OCR views in random order (Random).

be realistic, depending on the domain, to expect a high quality in-domain text corpus. For example, there might be a corpus of already electronically available documents. Therefore we compare off-the-shelf LMs — which were pre-trained only on a standard web corpus — and LMs that we further adapted on high-quality in-domain text. We use 60% of the data in each case for the domain adaptation step.

Reported settings. For the RETAS dataset we generated 3 settings, as each book can contain more than one OCR view: (i) combining the two best views, (ii) the best and the worst, and (iii) iteratively combining all in a random order. For ICDAR we only have 2 views. We analyze three scenarios: (i) numerical data, (ii) non-numerical data, and (iii) full data. We report the WER absolute results in Tab. 1, and the improvements achieved by our system on RETAS and ICDAR in Tab. 2 and Tab. 3 respectively. The numbers indicate the improvement over both the best (b) possible input view and over the worst (w) input view.

5.3 Results

RETAS. Tab. 2 shows that without domain adaptation, when the two best views are combined (*Best+2nd*), results are positive with an improvement of up to 8% with respect to the best view. For *Best+Worst*, the results deteriorate with respect to the best view, but always improve with respect to the worst. The results for *Random*, in which all 96 available views are iteratively combined in a random order, the results are worse than *Best+Worst*. This shows that this method would not automatically yield a good result without some prior selection of good OCR systems.

Model	Numeric		Non-Numeric		All	
	b	w	b	w	b	w
without domain adaptation						
3-gram	-23%	43%	-21%	46%	-23%	44%
GPT	-15%	47%	-24%	45%	-18%	46%
GPT2	5%	56%	-5%	53%	2%	55%
GPT2XL	6%	57%	-2%	54%	4%	56%
with domain adaptation						
GPT	-2%	53%	4%	57%	0%	54%
GPT2	15%	61%	8%	59%	13%	60%

Table 3: ICDAR. WER relative improvement over best OCR view (b) and worst view (w). Results for only Numeric characters, Non-numeric characters and All.

Domain adaptation did not improve here, and there is even a general and slight deterioration of the results. This is likely caused by the fact that as the model was pre-trained on the same domain, it overfits some books without specific domain gains.

ICDAR 2019. This dataset is particularly challenging due to the quality of the images which leads to noisy OCR views. Also, unlike the text on RETAS, the context is organized in a receipt layout structure, which does not necessarily fit the autoregressive generation assumption of the LMs. This dataset is significantly more difficult for OCR systems, i.e., compared to the RETAS dataset Tab. 1 shows an absolute WER more than 12 times larger for the best OCR and almost 14 times larger for the worst OCR views. This means that the gains of any correction are very impactful.

Table 3 shows that only the GPT2 model with domain adaptation significantly outperforms the rest over the best OCR view for numeric data. Non-numerical data seems to be more challenging with only domain adaptation settings generating an improvement, we conjecture that this is due to sequences of symbolic characters that are common in the receipts. As in the RETAS dataset GPT2 achieves the best performance

5.4 Error Analysis

Our method tends to make mistakes when the errors occur at the very beginning of the sequence, propagating them as the sequence advances. Looking

at longer sequences via beam search has mitigated this to a certain extent (although the computational cost limits the lookahead). We believe that this is caused by the auto-regressive nature of the GPT models, which suggests that it might make sense to explore the use of bidirectional models. This would imply a set of challenges outside the scope of this work, such as a mechanism to score the sequence or a way to deal with OCR views with a different number of tokens.

The proposed approach also tends to fail when the OCR quality of one of the underlying systems is poor. This can be seen in Table 2: the experiments with random OCR views perform worse than the setting with two views. Such an issue is also present in the ICDAR dataset with just two views: whenever one of the views has significantly inferior quality, the reconciliation can be worse than the best view. This happens in around 8% of the test sequences and indicates that the quality of the underlying OCRs is still relevant. Thus, a way to estimate the quality of the OCR views in advance would be beneficial.

6 Conclusions and Future Work

We presented an approach for post-OCR corrections in an unsupervised way, relying on multiple OCR inputs and LMs for reconciliation. Our results show that the approach consistently improves over the single best input. We also show that in a dataset with a different domain with respect to the pretraining data, a domain adaptation step is able to significantly improve the performance. Questions that are not addressed in this study and which are open for future work are: (i) how can one deliberately generate different OCR views in our setting; (ii) if there is also an unsupervised way to pick good OCR views, because as the *Random* setting on RETAS shows that just randomly merging views does not automatically yield the best possible result; (iii) whether it is possible to use a bidirectional LM; (iv) whether it is possible to assess in advance the quality of the underlying OCRs.

References

Chantal Amrhein and Simon Clematide. 2018. Supervised OCR error detection and correction using statistical and neural machine translation methods. *J. Lang. Technol. Comput. Linguistics*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of ICLR*.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2017. ICDAR2017 competition on post-ocr text correction. In *Proceedings of ICDAR*.

Rui Dong and David Smith. Multi-input attention for unsupervised OCR correction. In *Proceedings of ACL*, pages 2363–2372.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *Proceedings of ICDAR*, pages 1516–1520.

A. Kumar, Piyush Makhija, and Anuj Gupta. 2020. Noisy text data: Achilles' heel of bert. In *W-NUT@EMNLP*.

D. Lopresti and J. Zhou. 1997. Using consensus sequence voting to correct ocr errors. *Computer Vision and Image Understanding*, 67:39–47.

William B. Lund, Douglas J. Kennard, and Eric K. Ringger. 2013. Combining multiple thresholding binarization values to improve OCR output. In *Proceedings of DRR*.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural OCR post-hoc correction of historical corpora. *CoRR*.

Eugene W. Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(2):251–266.

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 JCDL*, page 333–336.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Caitlin Richter, Matthew Wickes, Deniz Beser, and Mitch Marcus. 2018. Low-resource post processing of noisy OCR output for historical corpus digitisation. In *Proceedings of LREC*.

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2019. ICDAR 2019 competition on post-ocr text correction. In *Proceedings of ICDAR*, pages 1588–1593.

Stuart Russell and Peter Norvig. 2002. *Artificial intelligence: a modern approach*. Pearson.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of EMNLP*, pages 2716–2726.

Benno Stein, Dennis Hoppe, and Tim Gollub. 2012. The impact of spelling errors on patent search. In *Proceedings of EACL*, page 570–579.

Shaobin Xu and David Smith. 2017. Retrieving and combining repeated passages to improve OCR. In *Proceedings of JCDL*, pages 1–4.

Ismet Zeki Yalniz and Raghavan Manmatha. 2011. A fast alignment scheme for automatic OCR evaluation of books. In *Proceedings of ICDAR*, pages 754–758.

Takafumi Yamazoe, Minoru Etoh, Takeshi Yoshimura, and Kousuke Tsujino. 2011. Hypothesis preservation approach to scene text recognition with weighted finite-state transducer. In *Proceedings of ICDAR*, pages 359–363.