Lights, Camera, Action!
A Framework to Improve NLP Accuracy over OCR documents

Amit Gupte*
Alexey Romanov*
Sahitya Mantravadi*
{amgupte,samantr,alromano}@microsoft.com
MAIDAP, Microsoft
Cambridge, MA, USA

Dalitso Banda*
Jianjie Liu*
{jianjl,dabanda}@microsoft.com
MAIDAP, Microsoft
Cambridge, MA, USA

Raza Khan[†]
razaurrehman@gmail.com
Azure AI, Microsoft
Bellevue, WA, USA

Lakshmanan Ramu Meenal
laramume@microsoft.com
Azure AI, Microsoft
Bellevue, WA, USA

Benjamin Han
diha@microsoft.com
Azure AI, Microsoft
Bellevue, WA, USA

Soundar Srinivasan
sosrini@microsoft.com
MAIDAP, Microsoft
Cambridge, MA, USA

ABSTRACT

Document digitization is essential for the digital transformation of our societies, yet a crucial step in the process, Optical Character Recognition (OCR), is still not perfect. Even commercial OCR systems can produce questionable output depending on the fidelity of the scanned documents. In this paper, we demonstrate an effective framework for mitigating OCR errors for any downstream NLP task, using Named Entity Recognition (NER) as an example. We first address the data scarcity problem for model training by constructing a document synthesis pipeline, generating realistic but degraded data with NER labels. We measure the NER accuracy drop at various degradation levels and show that a text restoration model, trained on the degraded data, significantly closes the NER accuracy gaps caused by OCR errors, including on an out-of-domain dataset. For the benefit of the community, we have made the document synthesis pipeline available as an open-source project.

CCS CONCEPTS

• **Applied computing → Document management and text processing**; • **Computing methodologies → Information extraction**.

KEYWORDS

natural language processing, optical character recognition

1 INTRODUCTION

Despite years of advancement of information technologies, a vast amount of information is still locked inside *analog* documents. For example, the JFK Assassination Records Collection[1] consists of over 5 million pages of records. Document digitization technologies like Optical Character Recognition (OCR) have made it possible to index the collection at a *word* level, allowing search via keywords and phrases. More elaborate NLP technologies such as Named Entity Recognition (NER) can then be applied to extract information such as person names, allowing information retrieval at *entity* level. However, even modern OCR technologies can produce output of a poor quality depending on the quality of the scanned documents.

In this paper, we demonstrate an effective framework for mitigating the impact of OCR errors on any downstream NLP task, using the task of NER as an example. Although there is a growing body of work dedicated to OCR and quality improvements, the specific topic of the impact of OCR errors on NER has not been widely explored. With the exception of [7, 11], and to an extent [13], the majority of the previous work has focused on general OCR error detection and correction [4, 6]. Here, we focus on a framework that allows us to improve the accuracy of any downstream NLP task such as NER.

Our major contributions are as follows. (1) To address the scarcity of entity-labeled OCR documents for model training, we design a pipeline Genalog for generating analog synthetic documents. The pipeline takes plain texts optionally annotated with named entities, synthesizes degraded document images, runs OCR on the images, and propagates the labels onto the OCR output texts. These imperfect texts can then be aligned with their clean counterparts for model training. (2) We propose an *action prediction* model that can restore clean text from OCR output and mitigate the downstream NER accuracy degradation. (3) We systematically investigate NER accuracy drop on OCR output at various synthetic degradation levels, and show that our text restoration model can indeed significantly close the NER accuracy gap.

The core capabilities of Genalog – synthesizing visual document images, degrading images, extracting text from images, and performing text alignment with label propagation – can have a wide array of applications. To facilitate further research, we are making Genalog available as an open-source project on GitHub[2].

2 RELATED WORK

Since we focus on closing the accuracy gap of NER induced by OCR errors, we adopt a more conventional NER model and keep

---

it constant in our experimentation, instead of using the most recent transformer-based [24] NER model such as [25]. LSTM-CNN models [2] typically employ a token-level Bi-LSTM on top of a character-level CNN layer. Hence, we train a multi-task Bi-LSTM model that operates at both levels [25]. This provides more granularity at the character-level than the traditional Bi-LSTM CRF model [10].

To systematically study the challenges of performing NER on OCR output, [7] categorized four types of OCR degradations and examined the decrease in NER accuracy for each as evaluated on synthetic documents. Their analysis referenced DocCreator [14] as a tool for generating synthetic documents. Our implementation, Genalog, was initially inspired by this tool. Another work, [5], presented a flexible framework for synthetic document generation but did not produce annotations for a downstream task.

The work of [23] examined the impact of noise introduced by OCR on analog documents on several downstream NLP tasks, including NER. They observed a consistent relationship between decreased OCR quality and worse NER accuracy. [15] explored the relationship between the word error rates of noisy speech and OCR on downstream NER, and [19] noted the difficulty of extracting names from noisy OCR text.

The goal of our text restoration approach is to reconstruct a "clean" version of the text from OCR output, which may contain various misspellings and errors. Sequence-to-sequence (seq2seq) models are a natural first approach: they convert an input sequence into an output sequence and can process text at the character level [6]. The use of LSTMs [9] is a common paradigm for seq2seq models. Another approach is the use of an encoder-decoder model with a single or multiple LSTM layers [4, 22]. Here, the encoder-decoder model is not based on a seq2seq architecture; instead it is a straightforward decoding of a character at every time step corresponding to the input sequence. These approaches are powerful but have a few drawbacks that we will explore in details in Section 4.

3 GENERATION OF SYNTHETIC DOCUMENTS

While there are many publicly available annotated datasets for NER, there are few targeting OCR output, i.e., document images containing texts with entities marked manually. To address this data scarcity, we developed Genalog, a Python package to synthesize document images from text with customizable degradation. It can also run an OCR engine on the images and align the output with ground truth to propagate the NER labels. The final result consists of degraded document images, corresponding OCR text, and NER labels for the OCR text suitable for model training and testing (Figure 1). We now describe each of these steps in the remainder of this section.
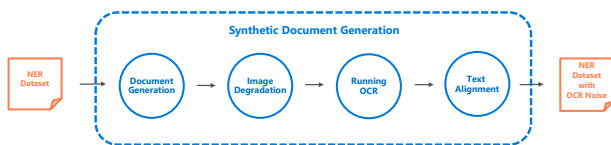

Figure 1: Synthetic document generation pipeline

3.1 Document Generation

A document contains various layout information such as font family/size, page layout, etc. Our goal is to generate a document image given a specified layout and input text. Genalog provides several standard document templates implemented in HTML/CSS (shown in Appendix A), and a browser engine is used to render document images. The layout can be reconfigured via CSS properties, and a template can be extended to include other content such as images and tables. In our experiments, we use a simple text-block template.

3.2 Image Degradation

Genalog supports elementary degradation operations such as Gaussian blur, bleed-through, salt and pepper, and morphological operations including open, close, dilate, and erode. Each effect can be applied at various strengths, and multiple effects can be stacked together to simulate realistic degradation. Figure 5 provides more details on each degradation.

3.3 Running OCR

We use a commercial OCR API to extract text from document images. Genalog calls the service on batches of documents and obtains extracted lines of text and their bounding boxes on each page. Genalog also computes metrics measuring the OCR accuracy, including Character Error Rate (CER), Word Error Rate (WER), and two additional classes of metrics: edit distance [16] and alignment gap metrics. This provides information on OCR errors and can capture errors such as **"room"** misrecognized as **"roorn"** by OCR.

3.4 Text Alignment

After obtaining the OCR text from synthetic documents, we propagate NER labels from the source text to the OCR text by aligning these texts and propagating the labels accordingly. Genalog uses Needleman-Wunsch algorithm [17] for text alignment on the character level. Because the algorithm is $O(n^2)$ in both time and space complexity, it can be inefficient on long documents. To improve efficiency, we use the Recursive Text Alignment Scheme [26] and search for unique words common to both ground truth and OCR text as anchor points to break documents into smaller fragments for faster alignment, thereby obtaining a 15-20x speed-up on documents with average length of 4000 characters.

3.5 Degradation Effect on OCR Error Rates

To understand what degradation on synthetic documents is realistic, we compute CER and WER on OCR texts obtained from a large corpus of *real* document scans (first row of Table 1). We then synthesize documents from the public CoNLL 2003 [21] and CoNLL 2012 [20] datasets with different degrees of degradation (none, light and heavy) and compute CER and WER on the OCR output. The result shows that in terms of WER, the *light* degradation is closer to the real degradation than the *heavy* degradation.

4 TEXT RESTORATION MODEL

To make our framework for mitigating OCR errors flexible, we propose a model for *text restoration*. This model can be trained

Lights, Camera, Action!
A Framework to Improve NLP Accuracy over OCR documents

Document Intelligence Workshop at KDD, 2021, Virtual Event

Dataset	Degradations	CER (%)	WER(%)
Real	-	1.2	7.3
CoNLL '03	None	0.3	2.5
CoNLL '12	None	0.3	2.5
CoNLL '03	All (light)	0.5	5.4
CoNLL '12	All (light)	0.9	7.5
CoNLL '03	All (heavy)	9.7	36.6
CoNLL '12	All (heavy)	8.5	33.4

Table 1: Error rates on Real and Synthetic OCR data

(a) Bottom: OCR output. Top: ground truth. Notice the growing character shift.

(b) Bottom: OCR output. Top: target characters to restore. Action prediction does not have character shift.
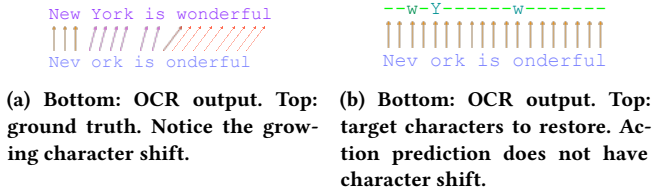
Figure 2: Problem of character shift and its mitigation.

(a) Distribution of actions

(b) Actions per character

Figure 3: Actions on CoNLL-2012, All Degradations

independently from the downstream task, and training data can be readily obtained from a data synthesis pipeline such as Genalog.

One approach to generate a corrected sequence of text is to correct one word at a time via a seq2seq model such as [6]. In that work they report when allowing an entire sentence as input instead of a single word, the model trained with smaller dataset resulted in *much increased* WER.

An alternative approach is to adopt an encoder-decoder model that decodes one character a time in sync with the input. Although this approach is faster than seq2seq, as insertions or deletions accumulate in a long sequence, the model needs to account for a growing character shift (see Figure 2a). This issue often leads to a repetition of the same character for several timesteps during the prediction. [4] mitigated this problem by limiting the length of the sequence to 20 characters and utilizing a sliding window. This significantly slows down inference and limits the context of the model.

Our solution builds on the previous approaches, but instead of predicting characters at each time step, we predict *actions* that are required for restoration (Figure 2b). For example, to restore the source text "Cute cat" from the OCR output "Cute at", we need to *insert* a "c" before the "a". This is akin to sequence labelling: each character in the input sentence is assigned an action labe). There are four main actions (INSERT, REPLACE, INSERT_SPACE, and DELETE) and two auxiliary actions (NONE and PAD). The first three actions need a character, e.g., we need to specify which character to INSERT. Predicting actions and characters separately can reduce vocabulary size and label sparsity, which can be severe for languages such as Chinese or Japanese. The distributions of the actions on CoNLL-2012 is presented in Figure 3a.

Since it is a sequence labelling problem, each input character is limited to have only one action. If a sentence is missing several characters in a row, our model is limited to recovering only one character. [18] and [1], who suggested similar approaches for grammatical error correction, mitigate this by applying the model up to 3 times. In OCR we've observed that errors mostly occur in
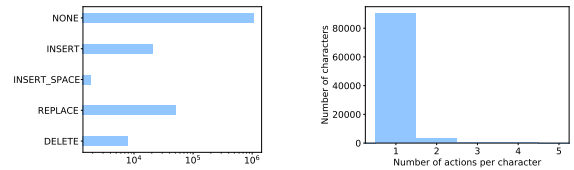
non-adjacent locations (see Figure 3b), and only few characters suffer from errors that need more than one action to fix. This is also supported by [12].After a manual examination of the characters that need exactly two actions, we found that most of them need the following two INSERT actions: insert a space and a character. We therefore introduce an action INSERT_SPACE that inserts both in one action.

Since OCR errors usually have single erroneous characters, fixing them does not require long contexts. Our model thus consists of a character embedding layer followed by one-dimensional convolution layers, and for each input character, it predicts an action and a character using two separate fully-connected layers. This model is trained with a weighted combination of cross-entropy losses for actions and characters: $\mathcal{L}_{total} = \alpha * \mathcal{L}_a + \beta * \mathcal{L}_c$.

5 EXPERIMENTS AND RESULTS

To investigate the effect of OCR errors on NER accuracy and the effectiveness of our model in mitigating them, we first use Genalog to create a synthetic dataset from clean texts ("clean data") and propagate the NER annotations to the OCR text ("noisy data"). We measure the the accuracy of the NER model on the clean data and the noisy data. Using the alignment between the ground truth and the OCR output, we produce labeled data to train the text restoration model. We then evaluate the same NER model on the restored text to determine if we can close the accuracy gap caused by OCR.

5.1 Data

We use the well-known corpora from CoNLL-2003 [21] and CoNLL-2012 [20] and refer to them as "clean data". Next, Genalog synthesizes and generates equivalent OCR texts, called "noisy data". For each degradation effect in Genalog, we generate three versions of noisy data: none, light, or heavy degradation. This allow us to observe, at a finer-grained level, the influence of degradations on NER accuracy. We also produce a version of datasets with *All Degradations*. Appendix B shows example images of each degradations and Appendix C lists degradation parameters.

5.2 NER Model

We use a popular character-level and token-level Bi-LSTM model [3] for NER. The model consists of a char-level Bi-LSTM layer followed by a token-level Bi-LSTM. Since we are using CoNLL-2012 and CoNLL-2003 to train the model, we utilize separate classification layer for each dataset [25]. We settled on this popular approach since our focus is mitigating NER accuracy drop caused by OCR, not achieving state-of-the-art NER on clean text.

5.3 Results and Analysis

Accuracy of the Action Prediction Model. Table 2 compares character-level and word-level accuracy between the noisy OCR text and the restored text. We observe that our model improves accuracy according to both metrics, and small improvements at character level lead to bigger improvements at word level.

Dataset	CoNLL-2012		CoNLL-2003	
Degradations	Light	Heavy	Light	Heavy
Char Accuracy Noisy	0.986	0.900	0.989	0.900
Char Accuracy Restored	0.991	0.907	0.994	0.903
Word Accuracy Noisy	0.927	0.661	0.942	0.646
Word Accuracy Restored	0.962	0.732	0.969	0.700

Table 2: Character-level and word-level accuracy of the noisy text and restored text

There are several common OCR errors that our model is able to recover, including errors introduced by line breaks and hyphenation (e.g., "rec-ognize" to "recognize"). It also restores words suffering from multi-character errors. Table 13 shows more examples.

NER Accuracy on Noisy Text. To improve NER accuracy on noisy text, we first pass it through the action prediction model trained on both CoNLL-2003 and CoNLL-2012. The restored text is then fed to the NER model to generate entities. To measure the benefit in NER accuracy due to the use of restoration, we first examine the drop in NER accuracy introduced by degradation and OCR. We then evaluate NER on the restored text. By comparing the NER accuracy on the noisy text to that on the restored text, we can then compute the improvement introduced by the restoration model. Table 3 shows the result of this evaluation using All Degradations (Light) on CoNLL-2012 and CoNLL-2003. The NER accuracy on noisy text goes down significantly on all datasets. Our restoration approach is able to improve the F1 scores considerably, and the accuracy gaps are reduced up to 73% on CoNLL datasets.

Dataset	CoNLL-2012	CoNLL-2003	CNN
Clean Text	0.832	0.860	0.989
Noisy Text	0.783	0.820	0.590
Restored Text	0.819	0.841	0.895
Relative Gap Reduction	73%	52%	76%

Table 3: NER F1 score on the clean, noisy, and restored text. See subsection 5.3 for the description of the CNN column.

Degradation	None	Bleed	Blur	Pepper	Phantom	Salt	All
Noisy Text	0.807	0.561	0.780	0.742	0.795	0.749	0.517
Restored Text	0.828	0.656	0.813	0.785	0.824	0.798	0.576
Gap Reduction	84%	35%	64%	47%	77%	59%	19%

Table 4: NER F1 score on the noisy and the restored text from the CoNLL-2012 test set with heavy degradation.

To further investigate the effect of *heavy* degradations, we applied these degrations to CoNLL-2012 dataset and measured the resulting errors in NER. Table 4 shows that all degradations hurt the NER accuracy and, in all cases, the proposed restoration model helps to improve the NER scores. The NER accuracy on degraded text declines the most when all degradations are combined, but 19% of this gap is recovered by our model. In general, we notice that the action prediction model works the best at moderate levels of degradation, where up to 77% of the accuracy drop is restored. When the text is obscured by many degradations in the heavy mode, the gap reduction becomes smaller.

NER Accuracy on Out-of-domain Datasets. To evaluate how our model generalizes to unseen documents from another domain, we evaluated our framework on a new[3] "CNN" dataset – 1000 randomly sampled articles from the CNN part of the DeepMind Q&A Dataset [8]. Since the CNN datset does not have NER labels, we run our NER model on it to produce the "ground truth". We then use `Genalog` to generate synthetic documents with `All Degradations (Light)`.

The accuracy of the NER system on the degraded text drops from the perfect score (the ground truth labels are produced by the same system)[4] to the F1 score of 0.59. Notice that the NER accuracy gap on the CNN dataset is much bigger than that on the CoNLL datasets (40 vs. 4.5 points), because here we are evaluating the accuracy using the model output on clean text as ground truth. Any change in NER prediction due to OCR errors therefore will contribute to the accuracy gap, while in the case of CoNLL datasets, changing an *incorrect* NER prediction (based on human annotations) to another incorrect prediction does not contribute to the accuracy gap.

After applying the restoration model to the degraded text, the accuracy of the NER system rises to 0.895 F1. Despite not being trained on the CNN dataset, our action prediction model is still able to close 76% of the NER accuracy gap. This result shows the action prediction model is able to generalize to data from another domain.

6 CONCLUSION

In this paper, we have demonstrated an effective framework to mitigate errors produced by OCR, a key step in the process of document digitization. We first constructed a data synthesis pipeline, `Genalog`, capable of generating synthetic document images given plain text input, running the images through OCR, and propagating annotated NER labels from the input text to the OCR output. We then proposed a novel approach, the *action prediction model*, to restore text from the inflicted OCR errors and show that our model does not suffer from problems faced by the conventional models when alignment mismatches between input and output accumulate. Lastly, we demonstrated that the accuracy of an important downstream task, NER, does drop at various degradation levels, and our text restoration model can significantly close the accuracy gaps, including on an out-of-domain dataset. Since the design of our restoration model does not depend on a the downstream task, it is generalizable to many other NLP tasks as well.

---

[3] Although CoNLL includes news articles, they are just a part of it.
[4] It is less than 1.0 due to sentence splitting differences between label generation and evaluation.

Lights, Camera, Action!
A Framework to Improve NLP Accuracy over OCR documents

Document Intelligence Workshop at KDD, 2021, Virtual Event

REFERENCES

[1] Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel Iterative Edit Models for Local Sequence Transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4251–4261.

[2] Jason Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4, 0 (2016), 357–370. https://transacl.org/ojs/index.php/tacl/article/view/792

[3] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)* (2017).

[4] Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2017. Generating a training corpus for OCR post-correction using encoder-decoder model. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1006–1014.

[5] D. Etter, S. Rawls, C. Carpenter, and G. Sell. 2019. A Synthetic Recipe for OCR. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 864–869.

[6] Kai Hakala, Aleksi Vesanto, Niko Miekka, Tapio Salakoski, and Filip Ginter. 2019. Leveraging Text Repetitions and Denoising Autoencoders in OCR Post-correction. *arXiv preprint arXiv:1906.10907* (2019).

[7] Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2019. An Analysis of the Performance of Named Entity Recognition over OCRed Documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 333–334.

[8] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*. 1693–1701.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[10] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[11] Wonseok Hwang, Seonghyeon Kim, Minjoon Seo, Jinyeong Yim, Seunghyun Park, Sungrae Park, Junyeop Lee, Bado Lee, and Hwalsuk Lee. 2019. Post-{OCR} parsing: building simple and robust parser via {BIO} tagging. In *Workshop on Document Intelligence at NeurIPS 2019*. https://openreview.net/forum?id=SJgjf695UB

[12] Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, Antoine Doucet, et al. 2019. Deep statistical analysis of OCR errors for effective post-OCR processing. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 29–38.

[13] Axel Jean-Caurant, Nouredine Tamani, Vincent Courboulay, and Jean-Christophe Burie. 2017. Lexicographical-Based Order for Post-OCR Correction of Named Entities. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. IEEE, 1192–1197.

[14] Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. 2017. Doccreator: A new software for creating synthetic ground-truthed document images. *Journal of imaging* 3, 4 (2017), 62.

[15] David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 2000. Named entity extraction from noisy input: speech and OCR. In *Sixth Applied Natural Language Processing Conference*. 316–324.

[16] Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press.

[17] SB Needleman and CD Wunsch. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology* (1970).

[18] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR–Grammatical Error Correction: Tag, Not Rewrite. *arXiv preprint arXiv:2005.12592* (2020).

[19] Thomas L Packer, Joshua F Lutes, Aaron P Stewart, David W Embley, Eric K Ringger, Kevin D Seppi, and Lee S Jensen. 2010. Extracting person names from diverse and noisy OCR text. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. ACM, 19–26.

[20] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*. Association for Computational Linguistics, 1–40.

[21] Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147.

[22] Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. 2020. Optimizing the neural network training for OCR error correction of historical Hebrew texts. *iConference 2020 Proceedings* (2020).

[23] Daniel van Strien., Kaspar Beelen., Mariona Coll Ardanuy., Kasra Hosseini., Barbara McGillivray., and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH,*. INSTICC, SciTePress, 484–496. https://doi.org/10.5220/0009169004840496

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[25] Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics* 35, 10 (2019), 1745–1752.

[26] Ismet Zeki Yalniz and Raghavan Manmatha. 2011. A fast alignment scheme for automatic ocr evaluation of books. In *2011 International Conference on Document Analysis and Recognition*. IEEE, 754–758.

A DOCUMENT GENERATION

Genalog provides three standard document templates for document generation, depicted on Figure 4.

B DOCUMENT DEGRADATION

We show example degradations produced by Genalog in Figure 5.

C DEGRADATION PARAMETERS OF USED DATASETS

D ACTION PREDICTION MODEL ARCHITECTURE

Figure 6 provides an overview of the architecture of the action prediction model.

E EXAMPLE OF RESTORED SENTENCES

Table 13 shows examples of restored sentences together with OCR output and the ground truth clean data.

(a) Multi-column Document Template   **(b) Letter-like Document Template**   **(c) Simple Text Block Template**

Figure 4: The three standard document templates available in `Genalog`

Degradation	No Degradation	Bleed Heavy	Blur Heavy	Close Heavy	Open Heavy	Pepper Heavy	Salt Heavy	All Heavy
Word Accuracy	0.954	0.735	0.902	0.831	0.940	0.879	0.905	0.697
Character Accuracy	0.991	0.945	0.980	0.966	0.988	0.975	0.984	0.915

Table 5: OCR accuracy on degraded text, obtained from the CoNLL-2012 train set. Examples of the degraded documents are shown in Figure 5.

Order	Degradation	Parameter Name	Parameter Value	Description
1	open	kernel shape	(9,9)	Thickens characters
	-	kernel type	plus	Kernel filled with "1"s in a "+" shape
2	close	kernel shape	(9,1)	Remove horizontal structures
	-	kernel type	ones	Kernel filled all with "1"s
3	salt	amount	0.5 / 0.7	Substantially reduce effect with more salt
4	overlay	src	ORIGINAL_STATE	Overlay current image on the original
	-	background	CURRENT_STATE	
5	bleed through	src	CURRENT_STATE	Apply bleed-through effect
	-	background	ORIGINAL_STATE	
	-	alpha	0.8 / 0.8	Transparency factor
	-	offset x	-6 / -5	Offset (pixels) of the background in x-axis
	-	offset y	-12 / -5	Offset (pixels) of the background in y-axis
6	pepper	amount	0.015 / 0.005	Imitate page degradation
7	blur	radius	11 / 3	Apply Gaussian blur
8	salt	amount	0.15	Add digital noise

Table 6: The sequence, in order, of types of degradation, along with the associated parameter values that was utilized to produce the All Degradations (Heavy / Light) dataset. Please see Figure 5c and Figure 5b for example images. Note ORIGINAL_STATE refers to the original state of an image before any degradation, and CURRENT_STATE refers to the current state of an image after the last degradation operation.

Lights, Camera, Action!
A Framework to Improve NLP Accuracy over OCR documents

Document Intelligence Workshop at KDD, 2021, Virtual Event

(a) No Degradation

(b) All Degradations (Light)

(c) All Degradations (Heavy)

(d) Bleed-through (Heavy)

(e) Blur (Heavy)

(f) Open (Heavy)

(g) Bleed-through (Light)

(h) Blur (Light)

(i) Open (Light)

(j) Close (Heavy)

(k) Pepper (Heavy)

(l) Salt (Heavy)

(m) Close (Light)

(n) Pepper (Light)

(o) Salt (Light)

Figure 5: Example of degradations produced by `Genalog`

R   W

Action fully connected

Char fully connected

1d convolution

Char embedding

N e v   Y o k   i s

Figure 6: Action Prediction Model

Lights, Camera, Action!
A Framework to Improve NLP Accuracy over OCR documents

Document Intelligence Workshop at KDD, 2021, Virtual Event

OCR Sentence	Reconstructed Sentence	Ground Truth Sentence
No one will be able to rec- ognize her body !	No one will be able to recognize her body .	No one will be able to recognize her body !
Although ! still have n't made the final decision to go	Although I still have n't made the final decision to go	Although I still have n't made the final decision to go
Why do you think the North Koreans chose july Fourth /?	Why do you think the North Koreans chose July Fourth /?	Why do you think the North Koreans chose July Fourth /?
israel says & may abandon the peace negotiations altogether .	Israel says it may abandon the peace negotiations altogether .	Israel says it may abandon the peace negotiations altogether .
! will protect this cky and save it .	I will protect this city and save it .	I will protect this city and save it .

Table 13: Sample of sentences that were correctly restored by the action prediction model