# **Optimizing OCR Error Correction of Historical Newspapers in Hebrew using Neural Networks and Crowdsourcing /** Omri Suissa

## Abstract

In the last few decades, paper-based documents such as books and newspapers are digitized using digitalization technology called OCR (optical character recognition). These resources are essential both for research and preservation of cultural heritage. In numerous digital humanities projects, there is a need to analyze paper-based documents automatically. The first step towards this goal is to use an OCR to digitize paper-based documents. Unfortunately, OCRed historical texts still contain a significant percentage of errors that undermine further analysis and preservation. Neural networks have shown great success in solving Natural Language Processing (NLP) tasks, including spell checking. However, neural network training requires a vast amount of training data (pairs of input and output sentences) that does not exist in Hebrew. This is one of the reasons that in Hebrew there is limited research and no optimal neural network structure for fixing OCR errors.

This research examines how to optimize neural networks learning process for OCR error correction in Hebrew historical newspapers. To achieve this goal three aspects were tested: creating training dataset using crowdsourcing, optimizing the neural network structure for the OCR error correction task and the ability of the neural network to generalize for different content domains. Finally, a comparison between different OCR error correction algorithms (humans, industry-leading spell checks and neural networks) was made.

To optimize dataset creation, a crowd-sourcing experiment was launched using Amazon's Mechanical Turk[1] (AMT). This experiment tested the unique aspects of OCR post correction (methodology, scanned image and text length); while acknowledging this domain's limitations such as lack of golden standard, OCR unique spelling errors and segmentation errors. Every dataset item included an “OCRed” texts (with common OCR errors), “scanned image” and a gold standard

[1] https://www.mturk.com/

Library of Information Science
Bar-Ilan University, Ramat-Gan, Israel
Email: Ruthi.Tshop@biu.ac.il

הספריה למדעי המידע
אוניברסיטת בר-אילן, רמת גן 5290002
טל' 972-3-5318163 Tel.

(correct) text. The experiment tested the effect of the methodology of the proofing process, the effect of the scanned image and the effect of the text's length. Three measurements were developed to assess three strategies: quality, time efficiency, and effectiveness. A total of 753 crowd-workers fixed 3796 texts using AMT platform and a dedicated site that was built for the experiment. Using quality measurement, we found that when segmentation allows it, medium length texts are the most efficient length to fix (comparing to long or short texts) and should be fixed using a sub-tasks methodology (Find-Fix) accompanied with the scanned image. Surprisingly, when segmentation prevents splitting long text, we found that the straightforward (naive) one-shot proofing achieves better quality. Moreover, long texts and one-shot proofing
methodology always wins when it comes down to time efficiency. Finally, when balancing between
quality and time efficiency (effectiveness), we found that short texts are the most cost-effective should be used with a sub-tasks' methodology (Find-Fix). When segmentation prevents from splitting into short texts, the strategy becomes more complex. This paper results help reduce the complexity of the crowdsourcing strategy choice and have important practical implications for many digital humanities projects which aim to analyze the content of OCRed document
collections.

Optimizing the neural network's structure was done by comparing several aspects that influence the network's accuracy: depth, regulation (dropout), number of features, layer type, dataset size (epoch size), batch size and bidirectional learning. A series of experiments was conducted on an artificial dataset that was generated from the Ben Yehuda Project[2] (the Hebrew equivalent to Gutenberg[3] project). The optimized network improved the accuracy (compared to the baseline network) by 3% (from 85% to 88%) and the end result precision by 4%. The optimized network was based on four bidirectional LSTM layers with 500 hidden size, 20% dropout, 250,000 examples in every epoch and 256 examples in every batch.

Using the strategies developed in the crowd-sourcing experiment, another experiment was

[2] https://bybe.benyehuda.org/
[3] https://www.gutenberg.org/

launched using articles from the most extensive historical Hebrew newspapers collection - JPress[4]. Seventy-five students corrected 150 articles. Using this dataset, a Hebrew specific OCR error injection algorithm was designed, and another artificial dataset was created based on Ben Yehuda Project texts. The optimized network was trained on this dataset as well. To assess the network's sensitivity to the content domain, the network was trained on a Bible-based dataset (using the same error injection algorithm). Both datasets were tested with different noise ratios. These experiments show that the biblical language is too far from the modern Hebrew language to allow the network to learn and correct modern OCR errors. In addition, high noise ratio prevented the network from learning, and a very little (10%) noise was needed to achieve optimal learning.

To conclude, a comparison between the different types of error correction algorithms (human subjects, leading spell checkers by Microsoft and Google and the optimized neural networks) was performed. Even with the latest advancements in deep learning, humans were able to generalize far better when fixing OCR errors in an unfamiliar content genre. Human subjects' quality (contribution) was 15 times greater than the neural network, even after the optimization of the structure and the datasets. However, comparing to industry-leading spell checkers, the neural network improved the quality (contribution) and the accuracy while the industry-leading spell checkers introduced more errors than corrections.

These results are another step towards creating automated tools for historical Hebrew OCR correction and toward historical cultural preservation. Researches can use these results to reduce the complexity when designing neural networks for the OCR error correction and to improve the OCR process itself.

The results of this research can be a starting point for other researches in the field of OCR post correction in Hebrew, spell checking in Hebrew (and other languages) and OCR accuracy improvement. Moreover, the optimization of crowdsourcing and neural network structure can be continued using aspects that were not tested in this research and by comparing to other content domains.

[4] http://web.nli.org.il/sites/JPress/Hebrew/Pages/default.aspx

006.424 סוי.או תשע"ט

9926564169405776

Library of Information Science
Bar-Ilan University, Ramat-Gan, Israel
Email: Ruthi.Tshop@biu.ac.il

הספריה למדעי המידע
אוניברסיטת בר-אילן, רמת גן 5290002
טל' 972-3-5318163 .Tel

www.is.biu.ac.il/library