October 2023

CCS Layout Analysis Benchmark

Introduction

For more than 20 years, CCS has been offering automatic layout analysis of newspapers as part of its software docWizz®. Initially, our layout analysis was designed as a complex image processing including an extensive set of rules applied on image level and to the results of the Optical Character Recognition (OCR).

In 2019 we integrated a Convolutional Neural Network (CNN) AI system to feed additional information to our rule-engine, allowing us to reduce the complexity of the rules and improve its robustness against variations in layout.

Recently we have replaced the AI system with a Detectron2 based network. This allowed us to replace the image processing and the rule engine with a generic post-processing module that is no longer specific to newspapers. Robustness is now controlled only by the training data.

Experiments with Transformer networks showed so far the same potential but at much higher computation cost.

Training Data

Based on contractual agreements with some of our clients, we are allowed to record corrections applied during manual quality assurance. We apply a mostly manual harmonization process to the data collected because quality standards in projects are usually below requirements for training data. Additionally, specifications vary between clients and projects. So far, we have produced an inventory of 175k pages with near-perfect layout analysis based on our harmonized specification. We identify 10 different types of zones: text-block, illustration, table, headline, advertisement, obituary, caption, running-title, author, and page-number.

In this case study we will focus on the detection of article headlines.

Headlines

From the technical perspective, headlines are the most important zones in a newspaper because they pave the way for the subsequent article segmentation. Headlines are also critical in supporting robust discovery on platform, with improved user experience through article-level indexing. Hence their automatic detection has a strong impact on productivity.

Evaluation

For evaluation, we compared three versions of our proprietary software. There was no 3rd party software that can provide such analysis.

docWizz 7.3, our last release without AI based layout analysis.

docWizz 8.0, our last version with CNN type of AI layout analysis.

docWizz 8.1, our recent release, the first with Transformer based layout analysis.

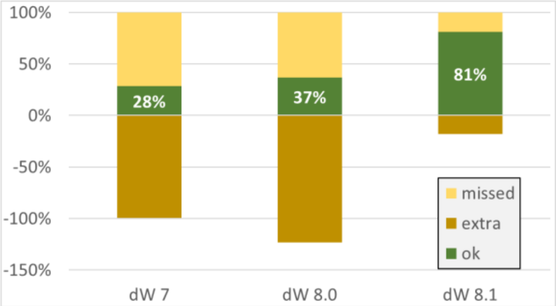
Ground Truth and measurement

Our set of ground truth comprises 175k pages of newspapers originating spanning the 19th, 20th and 21st centuries with 3-10 columns. The languages are mostly Latin alphabet with a small percentage of Cyrillic, Greek and Malay. Scans were made from microfilm and original prints.

From the ground truth we set apart a relatively small subset for testing and performed deep manual evaluation. We currently rely on manual evaluation as it allows developers to directly identify where to improve the training or what kind of training data should be added.

The test set was not used for training. Within the test set we counted the total headlines detected correctly (ok), the total number of headlines detected false positively (extra) and the headlines false negative headlines that are present in the ground truth but went undetected (missed).

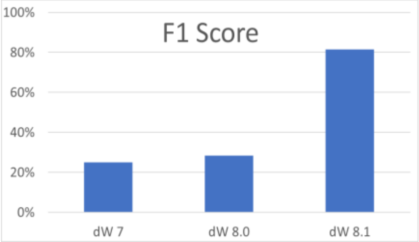
Results



Overall, the improvement achieved with dW 8.0 from dW 8.1 is substantial. The detection rate is massively increased while the false positive and false negative results are reduced substantially. Consequently, the F1 score reaches 81%. (F1 score is explained below)

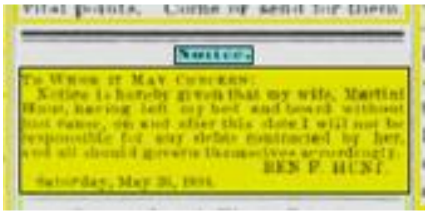
Care should be taken to extrapolate the results to other materials as the initial dataset for testing is relatively small.

Our evaluation indicate no correlation of the result to language, epoch, number of columns, or scanning source.



Typical errors

Extra headline errors produced by previous versions dW 7 and dW 8.0 were often small zones in advertisements or the title section wrongly identified as headline. With dW 8.1 the error schema has changed completely.



Vital points. Come or send for them.

Notice.

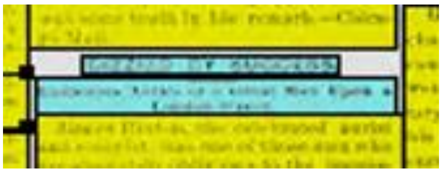
To Whom it May Concern:

Notice is hereby given that my wife, Martini Hunt, having left my bed and board without just cause, on and after this date I will not be responsible for any debts contracted by her, and all should govern themselves accordingly.

BEN F. HUNT

Saturday, May [illegible]

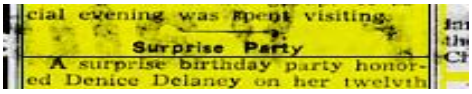
Now we see advertisements treated as articles (above) and errors resulting from inconsistencies in the training data (below).



[illegible]

Despite all efforts, our experts disagree on the correct zoning of the two lines.

Previous versions dW 7 and dW 8.0 missed many small headlines and wrongly classified them as text. This type of error can also be seen with dW 8.1 but with much lower frequency.



cial evening was spent visiting.

Surprise Party

A surprise birthday party honor-
ed Denice Delaney on her twelvth

A common theme of errors is that the model trained cannot learn contextual information. We do not feed the OCR results during training and the images are scaled down below readability.

We conclude that overall, the error profile of dW 8.1 closely resembles those of human operators.

Outlook

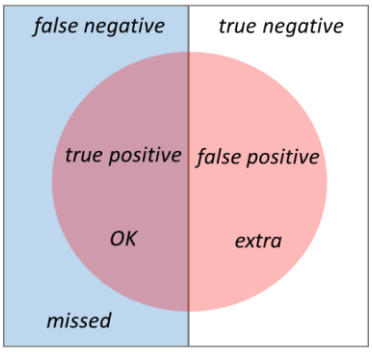
This evaluation will be updated and republished with further software updates and will include 3rd party evaluation solutions.

The use of automatic evaluation will allow an increase of the dataset used for testing and provide results that are statistically more reliable.

Because of the human-like error profile we plan to conduct an evaluation to compare fully automated zoning against human zoning in the context of Library of Congress' "National Digital Newspaper Program" (NDNP).

We believe the quality of automated article segmentation has achieved a level that is near equivalent to manual correction.

F1 score



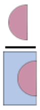
We apply the commonly used F1 score to measure performance. To Understand the scores, we use a graphical representation. Blue part of the square (left) contains all zones that are truly headlines, the white part (right) contains all other zones. The circle (red) contains all zones that are labeled as headlines by docWizz.

The F1 score first considers the terms "precision" and "recall". Precision is the answer to "How much of the result is true?"

$$precision = \frac{tp}{tp + fp} =$$

Recall is the answer to "How much of truth is in the result"

$$recall = \frac{tp}{tp + fn} =$$



F1 is then defined as:

$$F1 = \frac{2 * precision * recall}{precision + recall} = \frac{2tp}{2tp + fp + fn}$$

It is worth noticing that the F1 score is ignoring the true negative results. For a "needle in haystack" type of problem neither score seems adequate. However, for our layout analysis, we believe that the F1 score gives a very good indication of performance.