Improving the recognition of Dutch Gothic machine print, at four levels in the processing pipeline, in four days

Lambert Schomaker, Mahya Ameryan**,** Mirjam Cuper, Koen Dercksen, Jerry Guo, Rutget van Koert, Konstantin Todorov, Adriënne Mendrik & Xue Wang

Report on an *NWO/ICT with industry project* (5-day sabbatical/hackathon) focusing on OCR problems at the Dutch National Library (KB) and the Huygens Institute, January 20-24, 2020, Lorentz Center, University of Leiden

1. Problem statement

Libraries and archives are struggling with optical character recognition (OCR) of old machine-print fonts such as Gothic or 'fraktur'. This font was used in many important historical printed collections such as administrative texts and the then (17th century) newly invented 'newspapers' with interesting and detailed reports on important developments and events. When applying current state of the art OCR tools or sending the scanned images to large well-known companies that provide OCR services, the returned results are still quite disappointing. Problems are observed at all levels in the processing pipeline: binarisation suffering from ink bleed-through, layout analysis suffering from deviating page designs, marginalia and graphics, character recognition suffering from lack of pertinent font examples and font variation (Roman/Gothic) in a document and, finally, linguistic post processing suffering from an utter lack of encoded digital text corpora of suitable size. Actually, the OCR process is often intended to arrive at such corpora in the first place.

2. Approach

A team was formed to approach these problems in four days, with a fifth day for reporting (other teams were working on other industrial problems at the Lorentz Center, this week). The team decided to address problems at all levels in the processing pipeline.

3. Tasks

The processing pipeline consists of (1) image preprocessing, (2) layout analysis and segmentation into meaningful text objects (lines, words, characters), (3) text recognition and (4) linguistic post processing. From several possible scanned image collections, a suitable subset needed to be selected.

5. Data & research questions

The first day of the week was used to delve into the problem. Schomaker gave a crash course at the whiteboard on OCR to the participants who were from different research fields and literature hints were disseminated. The problem owners presented a host of data sets and problems. The post-processing group discovered an SQL database with recognized text and wanted to relate it back to the original images. In itself, this appeared not to be possible because the work flow in the application domain is more focused on the encoded digital text than on image (pixel) related information. However, inspection of the Alto .xml files from the

commercial OCR company that provided recognition results, it appeared that the bounding boxes (ROIs) of recognized strings were present in the .xml. This data set proved to be usable by all participants. It consists of a newspapers collection in Dutch Gothic script (1662 – 1795) from the Meertens Institute (Amsterdam), 15172 scans in Jpeg2000 format, which is common in libraries and archives, but not in science. The scans were converted to Jpeg by van Koert, who also provided a server with GPUs. Collaboration tools were Google Drive, Slack and Surf storage. The ensuing research questions (RQ) are:

1. Can we improve the recognition results by improving the binarisation of the scans?
2. Can we improve the detection of words, because the commercial OCR providers often produce very long strings in the recognized output, with no blanks between words
3. Given the problems at the character level, wouldn't it be better to train at the level of words, and how fast can we collect word labels from a tabula rasa setting, using high-performance computing and a human in the loop (the Monk approach)
4. Contrarily, can we use the suboptimal recognition results from a commercial provider (i.e., without human-validated ground truth) to train our own recognizers from scratch and allow many recognizers to be applied to the same data?
5. What are the effects of using state-of-the art text transduction tools to improve the quality of the OCR output to something that is closer to the intended language?
6. How does the performance evaluation work, at the different four stages in the processing pipeline?

6. Methods & Results

RQ1. Binarisation – The approach is to apply a number of traditional and deep-learned binarisation methods to a selected subset of scans from the Meertens newspaper data. Because no ground truth on the pixel intensities (ink/background) exist, as in the international DIBCO competition, another performance indicator is used, i.e., the final recognition performance of a common free OCR tool, Tesseract, on the data, which is used as the measuring device.

Jerry Guo demonstrated that binarisation methods are less important than envisaged. The traditional methods deteriorate the OCR results. Deep-learning (Deep Otsu) does not deteriorate the recognition performance noticeably, but leaving the input images – as is – gives the OCR performance that can be reached.

RQ2. Word segmentation – The HUC (problem owner) already had experience with using ARU-net (a U-net CNN, i.e. a deep-learning method) for baseline estimation. Because the Alto .xml output contains the ROI on the position of blanks via the '<SP>' tag, a new network can be trained, end to end, combining the line segmentation and word segmentation in an innovative single pixel-to-pixel transform.

Xue Wang showed that deep-learning based line segmentation (ARU-net based) can be enhanced by adding word-segmentation suggestions, in a pix2pix manner. The model was trained end-to-end, using recognition results (Alto.xml) of a commercial provider.

RQ3. Full-word recognition – The existing Monk e-Science service was used to bootstrap a

label collection process, starting from the tabula rasa condition), i.e., no labels. Scans were segmented into lines automatically and oversegmented into word candidates. Users start to label some correctly segmented words and an immediate process of data mining starts in the background using nearest-neighbour matching (upon the first label) and nearest-mean matching (with Nlabels > 1), in a high-performance computing setup. This is done using an alternation between recognition and retrieval, yielding ranked hit lists that can easily be labeled: the 'Fahrkunst' principle, Schomaker & van Oosten (2014).

Lambert Schomaker and Mirjam Cuper showed that the Gothic text could be ingested by the Monk system. Manual labeling with machine support yielded 319 word classes, 1760 human-labeled images in under two hours. This data can be used to evaluate different recognizers.

RQ4. Fresh, end-to-end training of a CNN/LSTM using the suboptimal recognition results of a commercial OCR provider as the training set. The OCR output is pruned to contain word-like strings, starting with a letter. A small ensemble of five classifiers using plurality voting and tie resolution is used. The performance evaluation is case insensitive.

Mahya Ameryan showed that an LSTM developed at AI@RuG could be trained end-to-end, achieving a word-recognition rate of 88%. Notably, the recognizer was trained on the output of a commercial recognizer, i.e., using word regions of interest and their OCR text as ground truth (100% can not be reached). The result is favorable and compares to a character recognition accuracy of 97.5% for a language with five-letter words, on average.

RQ5. A current state-of-the art approach to language translation is to use word embeddings (cf. Latent semantic indexing and the later word2vec approach) followed by an LSTM encoder and an LSTM decoder. For word embeddings, we use BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018), a pretrained network that is fine tuned on the text to be expected in the Meertens Newspapers corpus.

Koen Dercksen and Konstantin Todorov used the famous BERT (word-embedding method by Google) system to fine tune it (post train it) using the Meertens Newspaper texts, with promising results.

RQ6. How to provide a modular pipeline and work flow, such that users in the application domain can quickly evaluate different variants of processing? Ideally, tests are performed on the grounds of a neutral party, such as a national e-Science server. Is this possible in all cases, e.g. using virtualization or 'Docker' containers? In any case the overall concept for the framework can be designed in the short time frame of the current project.

Adriënne Mendrik modeled the work flow and processing pipeline in order to realize a modular, adaptable framework in which individual component variants can be inserted to assess their effect on evaluation. Performance metrics are considered within their particular usage context, e.g., transcription and retrieval require different performance metrics, and the object under consideration (characters vs words) also plays a central role in understanding the performance of an OCR system.

8. Conclusion

This was an exciting project week with tangible results, that are expected to have an impact on the 'industry', i.e., the KB and Humanities Cluster (HuC).
Notable is the fact that there is much more information in the recognition results of commercial providers than is currently used. Massive amounts of training data can be extracted. Apparently good recognition results can be achieved by alternative recognizers that are trained on such imperfect data. By using user-friendly labeling tools, benchmark data sets can be developed by the institutes themselves, which puts them in a more comfortable position in dealing with OCR providers. The results present a promising picture regarding the future integration of Gothic-font documents in current search engines such as 'Delpher'.

9. References

Ameryan, M. & Schomaker, L.R.B. (2019), A limited-size ensemble of homogeneous CNN/LSTMs for high-performance word classification, arXiv:1912.03223

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018).
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs.CL]

Gruening, T., Leifert, G., Strauß, T., Michael, J. & Labahn, R. (2019).
A Two-Stage Method for Text Line Detection in Historical Documents, arXiv:1802.03345 [cs.CV]

van Oosten, J.-P. & Schomaker, L.R.B. (2014). Separability versus prototypicality in handwritten word-image retrieval, Pattern Recognition, 47(3), pp. 1031-1038

Thanks to the team!