A BRIEF INTRODUCTION

Ocromore is a toolset to improve the character recognition accuracy by combining multiple OCR results. The character accuracy is especially crucial for automatic structuring and research purposes.

It was developed for the DFG-Project Aktienführer-Datenarchiv. The Aktienführer is a reference work published annually between 1956-1999 as print book comprising data for companies listed at stock exchanges in Germany.

One of the goal of this project is to process this data and store it in a structured manner in a database for companies listed at stock exchanges in Germany. To optimize the OCR workflow and the result, several tools were created and combined to a toolchain.

The produced images are then used to create multiple OCR results. To simplify the communication and guarantee a unified project structure a multiple OCR-engine interface „crass“ was designed. The character accuracy is especially crucial for automatic structuring and research purposes.

The postprocessing step of „ocromore“ is described in detail in the following section. The content of the combined result is further structured with „hocr-parser“.

The whole process can be declared as character-confidence-based wordwise multi-sequence alignment method.

WORKFLOW OR THREE FOR ONE

From the images, like the example left, initially different OCR-results are generated. This project uses the OCR-engines, Tesseracl, Ocropus and Abbyy Finereader.

It is necessary to produce results with word bounding boxes for the alignment and character confidence for the voting process. The results differ not only in single recognized characters but also in the arrangement of the lines. This leads to the first section, the alignment.

The alignment splits into three subsection: line, word and character. While the first two alignment steps are based on bounding boxes, the third takes a whole other turn. To match all lines and words to their pendant a mean value with some tolerance of horizontal and vertical bounding box values are calculated. Although the wordspaces are now mapped to one another, the content is not.

For the character level, a multi sequence alignment algorithm, called msa, was implemented. Msa is frequently applied to find matching subsequences in biological sequences. Using msa with very long sequences and sequences containing more variety than a few base pairs like natural language (UTF-8) can be prohibitively computationally expensive and/or can lead to unmatchable combinations. For that reason, splitting into wordwise fragments is quite important. The algorithm was additionally extended, that not only the same character but also optically similar get aligned. When matching the subsequences gaps are created, which are filled with so-called wildcards.

In the last section, the voting process, every matched character set is subjected to a decision process based on the confidence. Because each engine is weighted differently, rescaling the confidences in the first place is essential. It turned out that it is worth taking a close look, as some OCRs have special recognition strength. In the rescaling process letters, numbers, symbols and even special characters can be weighted differently.

The decision making process sums up all confidences and picks the highest one. If a wildcard is picked, it will be removed and the text left and right get concatenated.

The whole process can be declared as character-confidence-based wordwise multi-sequence alignment method.

RESULTS - TRIED AND TESTED

The gained results are evaluated against groundtruth (GT) data with a standard tool (ISRI Analytic Tools).

Our evaluated datasets showed a character accuracy of 84.08% for Abbyy, 98.67% for Ocropus, 98.79% for Tesseract and 99.19% for the combined result, which represents an accuracy increase of 0.49% compared to the best single result and an error reduction of 33%.

“Ocromore” and the other mentioned projects are written in Python and released under the patronage of the Mannheim University Library as open source projects in Github (see section “Open source projects”).

Access to the database and more information about the DFG-Project Aktienführer-Datenarchiv: https://digi bib.uni-mannheim.de/aktienfuhrer/.