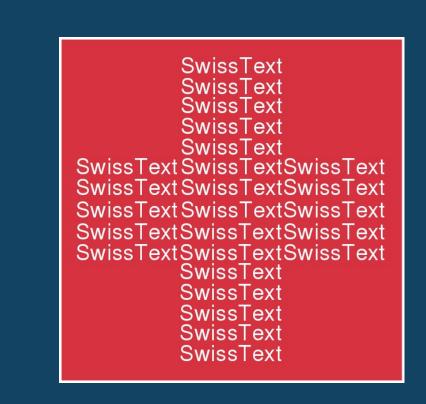# Understanding structural information in scanned documents

Holger Keibel

Karakun AG, Basel

SwissText

Abstract

HIBU is a proprietary solution platform based on which Karakun builds customer solutions around enterprise search and text analytics. In our customer projects we often deal with scanned documents which are digitized via OCR software. While for unstructured documents (running text), text analytics algorithms can directly operate on the OCR output as a subsequent and independent processing step, doing the same on structured sections of a document (such as forms and tables) tends to produce poor results because the OCR result does not capture the non-sequential relations in such structured sections (e.g., interpret a table cell relative to its column title).

This poster illustrates how this can be improved if the text analytics algorithms are in part integrated into the OCR step for these structured document sections. Example use cases from customer projects will be presented.

Introduction

Information extraction algorithms (like most NLP methods) typically process an input document as a sequence of words. This is well suited for running text. However, for more structured document sections, the layout may encode important semantic information, and this information is lost when the document is serialized into a sequence of words.

For instance, in a table, the content of each table cell needs to be interpreted relative to other table elements such as the corresponding column label, other cells in the corresponding row and often even the table caption. The information extraction mechanism needs to have access to this table structure in order to properly extract and interpret information from the table.

This poster addresses this challenge specifically for the case of scanned documents. Approaches exist that directly couple spatial/layout information with a specific extraction algorithm (Katti et al. 2018; Denk & Reisswig 2019). Our approach instead attempts to derive a logical representation of the overall table which can then be processed by any extraction algorithm.

Methodology

Input document: an excerpt of a land register (German)

Step 1: Segment document into textual and graphical sections

We apply LAREX (Reul et al., 2017). which in turn is based on the OpenCV library (see step 2). LAREX essentially detects graphical sections (such as images and tables, marked by red outline above) by pixel density and/or border lines.

Step 2: Extract table image
Cut at the outline coordinates to extract an image with just the table.

Step 3: Insert auxiliary lines on certain conditions

For this step, a first pass through steps 4-9 is executed.
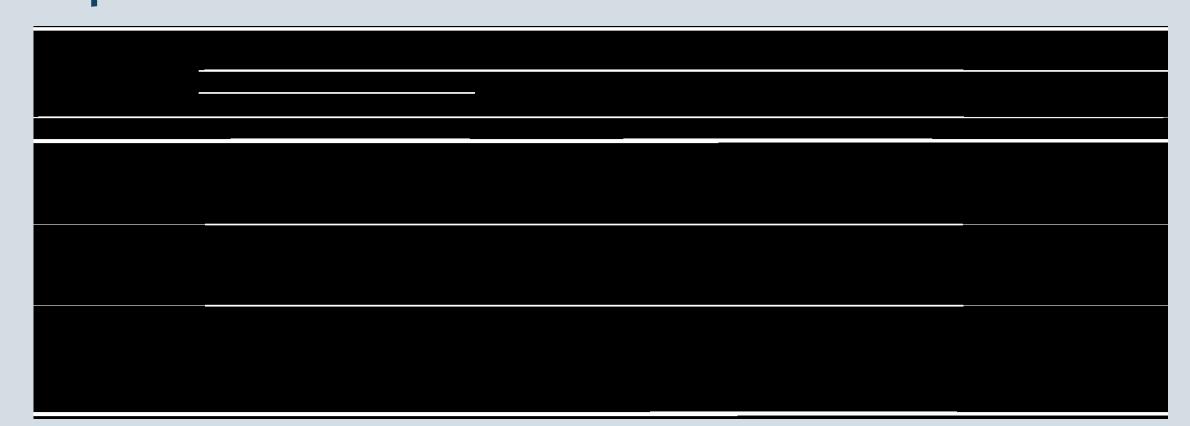
Step 4: Blur image to smoothen and repair lines
Steps 3 – 8 are all performed using OpenCV (https://opencv.org/, Apache 2 License).

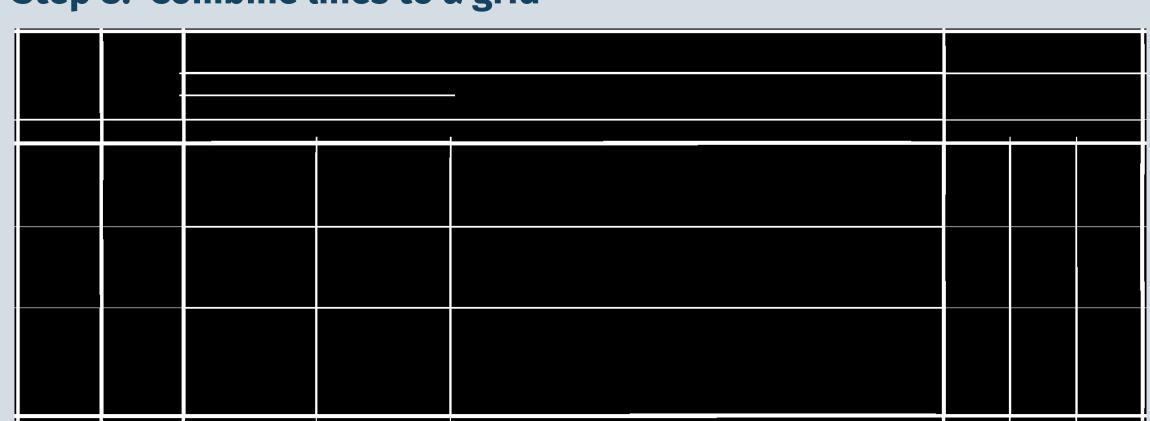### Step 5: Invert colors and binarize image to widen lines

Step 6: Derive horizontal lines and extend them

Horizontal lines are extended by means of dilation to ensure they will intersect with the vertical lines.

Step 7: In the same fashion derive vertical lines and extend them

Step 8: Combine lines to a grid

From the intersections, we can derive the rectangle coordinates of each table cell. Open rectangles are completed by implicit lines.

Step 9: Submit entire table to OCR engine and parse hOCR result
We extract from the hOCR result all non-empty XML elements "ocrx_word" and for each such element, use its coordinates to map it to the unique cell to which it belongs. Finally, for each cell, we combine all its "ocrx_word" elements in the natural order (left to right, top to bottom) to reconstruct the cell content.

Step 10: Resolve merged cells to derive structured representation

Postprocessing and use cases

The methodology described here is not fully generic but tuned to a particular type of table which encodes cell boundaries mostly by lines and partly by white space.

It is most useful for table formats with some degree of standardization, e.g. in documents issued by government agencies. For such cases, it is a very robust approach.

The resulting structured representation of OCR results can be submitted to any subsequent NLP chain applicable to the given problem.

For tables whose overall structure and semantics are known beforehand and whose cells are semantically simple, a rule-based approach generally suffices and does the job very reliably.

For tables with a more unknown structure or semantically more complex cells, an ML-based extraction method can take the structured representation as input and thereby substantially reduce the learning task.

Next steps

Integrate step 3 into step 9. Make logic less reliant on lines as row/column delimiters.

References

Denk, T. I & Reisswig, C. (2019). *Contextualized embedding for 2D document representation and understanding.* 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

Katti, A. R., Reisswig, C., Guder, C., Brard, S., Bickel, S., Höhne, J., & Faddoul, J. B. (2018). *Chargrid: Towards understanding 2D documents.* EMNLP 2018.

Reul, C., Springmann, U., & Puppe, F. (2017). *LAREX - A semi-automatic open-source tool for layout analysis and region extraction on early printed books.* In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage.

Acknowledgements