BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

TextBite: Segmentation of Logical Units in Text

Bc. Martin Kostelník*

Abstract

The goal of this project is the topic segmentation of text into coherent units. It builds on the PERO-OCR software, aiming to improve the processing of Czech historical documents and information retrieval for librarians and scientists. This included the creation and annotation of a custom dataset comprised of 4044 pages from books, dictionaries, and periodicals. We propose an innovative approach treating segmentation as a line clustering problem. Our method involves a two-stage process: initial detection of regions of interest containing text lines using the YOLOv8 model, followed by joining them using a graph neural network. This method achieves a V-measure of 75.59 %, 95.17 % and 89.32 % for books, dictionaries and periodicals, respectively.

*xkoste12@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

TextBite is a part of the semANT project at FIT VUT, which aims to enhance search capabilities in digitized documents by incorporating semantic understanding. TextBite focuses on the topical segmentation of historical documents into logical units like articles, dictionary entries, or news in a newspaper. We call these segments *bites*. This helps librarians and scientists who have large amounts of uncategorized data.

The project builds on the PERO-OCR software, which provides detection of text lines and their transcription. It does not however, guarantee the reading order. We therefore define the problem as a line clustering problem. Each cluster can be either a title or a text segment. The proposed methods are evaluated with clustering metrics: completeness, homogeneity, and V-measure.

2. Data

As no public datasets in the Czech language for topic segmentation were found, we decided to create our own. Three different types of historical documents are considered: books, dictionaries, and periodicals. In total, 4044 pages from historical documents were hand-picked from a digital library for processing. With the help of librarians and other students, these pages were annotated.

The most difficult and varied pages were then selected as validation and test datasets. Poster Section 1 showcases the distribution of pages and document types.

3. Proposed Method

We implemented a baseline solution to the problem without the use of machine learning techniques, which is based solely on the geometry of the page. This method finds a predecessor and successor of each line by geometric constraints. These relations might not be symmetric and if they are not, the connections are severed. The resulting segments are considered as bites for the baseline method.

Two improvements to the baseline method were created. The bites generated by baseline can be broken down further by introducing a heuristic function. The first heuristic is purely geometric. Separation is performed when the vertical distance of two consecutive text lines is greater than the average for the entire page, scaled by a hyperparameter.

The second heuristic is represented by a BERT-like language model fine-tuned for the next sentence prediction (NSP) task. Similarly to the distance-based heuristic, a separation is performed when the model predicts that two lines should now follow each other. A pre-trained CZERT [1] model was fine-tuned on the NSP task using our labeled data. We also pre-trained our own BERT-like models on a large corpus of book data from the digital library. Four models in varying

embedding sizes were trained using the traditional masked language modeling (MLM) and NSP tasks. They were then fine-tuned similarly to the CZERT model.

Contrasting with the baseline approach, the main proposed method involves a two-stage process: initial detection of regions of interest containing text lines using the YOLOv8 model, followed by joining them using a graph neural network. Pre-trained YOLOv8 models from Ultralytics [2] were fine-tuned to detect text regions in our data. The YOLOv8 nano, small, and medium variants were experimented with, also with varying image resolutions ranging from 640px to 1400px on the long side with the short side preserving the aspect ratio, and the best-performing model was selected. The regions are then matched with the text lines from PERO-OCR and can be considered bites.

However, YOLOv8 can only detect axis-aligned bounding boxes, and thus, bites spanning multiple columns cannot be detected. To overcome this, a complete, undirected graph is created from the detected regions and passed to a graph neural network. Regions represent the nodes and are composed of geometric features. The edges are represented by the euclidean distance of the two regions and most importantly, the cosine distance between CZERT embeddings of the text transcriptions. The effect of the network can be seen on the poster in Section 2.

The graph neural network serves as an edge classification model. It predicts whether two nodes (text regions) should be joined together. The residual gated graph convolutional layer [3] is used as its primary graph operator. This layer, along with dropout and layer normalization, forms a building block for the graph neural network. Multiple instances of these blocks are stacked to form the complete graph neural network.

The entire processing pipeline with the outputs of all methods can be seen on the poster in Section 4.

4. Results

The baseline method works fairly well with pages where each bite is one paragraph and the bites are separated by some distance. The distance-based heuristic further improves this, but the method struggles in dense and complex documents, especially in dictionaries, where the entries are not visually separated. The approach with language model improves the baseline considerably when CZERT model is used. Our own pre-trained models did not yield a reliable performance.

Even when considered as a standalone method, the YOLOv8 outperforms the baseline on all document types. The graph neural network further improves it by joining titles and bites spanning multiple columns. All numerical results can be seen on the poster in Section 3 along with the inference times. Examples of segmented pages can be seen on the poster in Section 5.

5. Conclusions

This project introduces a new method for text segmentation into logical units based on the YOLOv8 model and a graph neural network. A custom dataset of 4044 pages of historical documents was created and labeled. The method extracts the segments well even in documents with a complex layout.

The future matter in the project could involve further refining of all the processes. Image augmentation and a larger dataset would benefit the YOLOv8 detector. More text features using a better language model, like a topic embedding model, could improve the graph neural network.

Acknowledgements

I would like to thank my supervisor Ing. Karel Beneš for his expertise and guidance. Furthermore, I would like to express my thanks to Ing. Michal Hradiš Ph.D. for his valuable consultations.

References

[1] Jakub Sido, Ondrej Prazák, Pavel Pribán, Jan Pasek, Michal Seják, and Miloslav Konopík. Czert - czech bert-like model for language representation. *CoRR*, abs/2103.13031, 2021.

[2] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.

[3] Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *CoRR*, abs/1711.07553, 2017.