FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Supervisor's statement of a final thesis

Supervisor:	Ing. Milan Šulc, Ph.D.
Student:	Bc. Matyáš Skalický
Thesis title:	Relative Layout Matching for Document Data Extraction
Branch / specialization:	Knowledge Engineering
Created on:	25 May 2022

Evaluation criteria

1. Fulfillment of the assignment

▸ [1] **assignment fulfilled**
[2] assignment fulfilled with minor objections
[3] assignment fulfilled with major objections
[4] assignment not fulfilled

All points of the thesis assignment were fulfilled:
1. The review of the related work was done so thoroughly that it even lead to a research paper.
2. Matyáš prepared a benchmark for the task by implementing an efficient annotation tool and using it to label selected documents in the company.
3. The proposed baselines were implemented properly and showed to be very strong.
4. Matyáš proposed a method based on the recent advances in contrastive learning. He implemented the retrieval of similar documents and to match the corresponding areas.
5. The results were carefully evaluated and discussed.

2. Main written part   100 / 100 (A)

The thesis is well structured and easy to follow. The English level is good. All relevant sources are properly cited. I have to emphasize that Matyáš carried out the review of the related work in document information extraction so thoroughly it also contributed to a research paper (CLEF'2022, in review), where Matyáš is the first author. The thesis only contains the necessary content, and the CLEF paper is attached in the Appendix.

3. Non-written part, attachments   100 / 100 (A)

A considerable part of Matěj's work lied in the implementation and data preparation. Because of the absence of a dataset with layout annotations, Matyáš developed his own annotation tool to label the dataset efficiently and then used it to annotate the dataset as described in the thesis. Matyáš implemented the methods used in the thesis: the

"baselines" and the proposed method based on contrastive training. The implementation and debugging of the machine learning code were non-trivial and Matyáš showed good ML software engineering skills and habits, e.g. frequent design sanity checks (checking receptive field of the CNNs, gradients, etc.). The code is submitted as the supplementary material of the thesis. The dataset is proprietary and thus can not be published with the thesis, but it will be further used in the company.

4. Evaluation of results, publication outputs and awards   95 /100 (A)

The thesis review of related work was carried out so carefully it lead to a research paper. The experiments bring valuable insights beyond the end-to-end metrics, e.g. measuring the dependence of the accuracy on the number of similar documents in the dataset.
The baselines have shown to be very strong and the proposed method performs worse on most of the field types (with the exception of the total amount - the field "most typically moving with the length of the table"). While there are several directions in which the proposed method could be improved, in the expected scope of a master thesis the student did a great job. The negative results can not be considered a weakness of the thesis, and the student made sure to describe directions for potential improvements in the future work.

5. Activity of the student

▸ [1] **excellent activity**
[2] very good activity
[3] average activity
[4] weaker, but still sufficient activity
[5] insufficient activity

Matyáš was an active student, always prepared for our consultations, coming with his new ideas, results, and topics for discussion.

6. Self-reliance of the student

▸ [1] **excellent self-reliance**
[2] very good self-reliance
[3] average self-reliance
[4] weaker, but still sufficient self-reliance
[5] insufficient self-reliance

In the thesis, Matyáš proved his ability to independently work with the literature, propose methods based on the review of related work, and design and perform experiments.

The overall evaluation   100 /100 (A)

Overall, I rate Matyáš's master thesis as excellent: All points of the assignment were carried out carefully. The thesis is well structured and easy to follow, the English level is good. The thesis even lead to a research paper (submitted to CLEF'22). The assignment was complex and included a considerable amount of implementation work. Although the proposed method performed worse than the implemented baselines on most field types, the thesis brings valuable insights and proposes several directions for potential

improvements of the proposed method in the future work. Matyáš was an active student worked hard towards the thesis, it has been a pleasure working with him.

Instructions

Fulfillment of the assignment

Assess whether the submitted FT defines the objectives sufficiently and in line with the assignment; whether the objectives are formulated correctly and fulfilled sufficiently. In the comment, specify the points of the assignment that have not been met, assess the severity, impact, and, if appropriate, also the cause of the deficiencies. If the assignment differs substantially from the standards for the FT or if the student has developed the FT beyond the assignment, describe the way it got reflected on the quality of the assignment's fulfilment and the way it affected your final evaluation.

Main written part

Evaluate whether the extent of the FT is adequate to its content and scope: are all the parts of the FT contentful and necessary? Next, consider whether the submitted FT is actually correct – are there factual errors or inaccuracies?

Evaluate the logical structure of the FT, the thematic flow between chapters and whether the text is comprehensible to the reader. Assess whether the formal notations in the FT are used correctly. Assess the typographic and language aspects of the FT, follow the Dean's Directive No. 52/2021, Art. 3.

Evaluate whether the relevant sources are properly used, quoted and cited. Verify that all quotes are properly distinguished from the results achieved in the FT, thus, that the citation ethics has not been violated and that the citations are complete and in accordance with citation practices and standards. Finally, evaluate whether the software and other copyrighted works have been used in accordance with their license terms.

Non-written part, attachments

Depending on the nature of the FT, comment on the non-written part of the thesis. For example: SW work – the overall quality of the program. Is the technology used (from the development to deployment) suitable and adequate? HW – functional sample. Evaluate the technology and tools used. Research and experimental work – repeatability of the experiment.

Evaluation of results, publication outputs and awards

Depending on the nature of the thesis, estimate whether the thesis results could be deployed in practice; alternatively, evaluate whether the results of the FT extend the already published/known results or whether they bring in completely new findings.

Activity of the student

From your experience with the course of the work on the thesis and its outcome, review the student's activity while working on the thesis, his/her punctuality when meeting the deadlines and whether he/she consulted you as he/she went along and also, whether he/she was well prepared for these consultations.

Self-reliance of the student

From your experience with the course of the work on the thesis and its outcome, assess the student's ability to develop independent creative work.

The overall evaluation

Summarize which of the aspects of the FT affected your grading process the most. The overall grade does not need to be an arithmetic mean (or other value) calculated from the evaluation in the previous criteria. Generally, a well-fulfilled assignment is assessed by grade A.