Master Thesis Proposal: OCR post-processing of Swedish historical text using machine learning techniques

Dana Dannélls

2018-10-15

1. **Goal**: Train and test an OCR post-processing model that could be applied to Swedish historical text.

2. **Background**: Optical Character Recognition (OCR) is an automatic process for converting a digital image into a machine readable and editable text. The quality of OCR for modern print is very good but for older print accuracy varies substantially. Many errors are normally caused by irregularities in the print, the use of several typefaces, deterioration of the paper, and variation in the language.

One way to improve the quality of an OCR process is by applying statistical natural language post-processing methods, for example by combining a language and an error model, as follows:

\[
\arg\max_{\text{orig}} p(\text{orig}|\text{ocr}) = \arg\max_{\text{orig}} [p(\text{ocr}|\text{orig}) \cdot p(\text{orig})]
\]

where:

- error model = \( p(\text{ocr}|\text{orig}) \)
- language model = \( p(\text{orig}) \)

The error model estimates the probability that a certain transformation can occur to a string, and the language n-gram model defines the probability of the words in a document.

3. **Project description**: In the project, the student(s) should examine different machine learning techniques and try to improve the recognition accuracy of the OCR process for Swedish historical text.
4. **Recommended knowledge and skills:** The project requires knowledge of some programming language, and familiarity with machine learning techniques. Some knowledge of the Swedish language is a plus.

5. **Supervisors:** The work would be supervised by Dana Dannells, Språkbanken and possibly others from either Språkbanken or the Department of Computer Science and Engineering.

6. **References:**
