1 Project description

Currently, we are carrying out a project on the Swiss parliament archives. This corpus is formed by all the proceedings from 1891 to 1995, from both the National council and the council of states. From these documents we are studying how political interactions, interests, importance of socio-economic factors, etc., have been shaping the taking of political decisions throughout the years. Due to the number of documents and the ample time span, a project of this scale has never been carried out before, which underlines its potential.

Even though the scanned documents underwent a careful processes of digitization and optical character recognition (OCR) to extract the text, they still present ample errors, specially in early years. The nature of these errors is varied, and ranges from problems in the identification of texts’ bounding boxes, to a misalignment during the scanning process, or blurred characters. These lead to transcripts that are sometimes unreadable, hindering impossible their subsequent analysis with NLP methodologies.

In the present project we will tackle some of these problems, aiming at improving the quality of the extraction. The implemented methodology is not only intended to solve particular
problems of the parliament corpora, but rather is envisioned as a general methodology for the
analysis/discovery of errors and the correction of historical typewritten documents. As a first
approach to the intended pipeline, we could define the following steps. First, unsupervised or
weakly supervised methods may be used to identify specific errors in the pages, either coming
from the extracted text or the image of the scanned page itself. This will allow defining common
errors caused during the digitization. Once these errors are identified, in the second step we
will devise methods to correct them. We will mostly focus on transcription errors, as those are
expected to be the most common ones. But related errors, such as the wrong identification of
bounding boxes, might be also tackled. Finally, an interactive dashboard for the inspection of
the full process will be implemented. It will allow the user validating the detected errors and
the corrections provided.

Throughout this project, the candidate will have the opportunity of exploring different
deep learning methodologies, from the fields of computer vision (CV) and natural language
processing (NLP). By leveraging models such as Mask R-CNN [2] or YOLO [3], features of
wrongly detected bounding boxes could be identified. Transformers based language models such
as BERT [1] can help detecting sections of text with severe typos, as the computed embeddings
should capture syntactically wrong sentences. Through more traditional techniques, such as
measures of string similarity, errors at the level of independent words could be also captured.
Additionally, standard CV methods and image measures will allow identifying specific image
traits that could pinpoint to known-errors. Finally, for an improved OCR of the wrong text
sections, we will explore both open-source software that works directly on the original text
image, such as Tesseract [4], and ad-hoc models based on RNN architectures that could correct
directly the faulty transcript.

The present project will allow the student getting a larger familiarity with the usage of
different neural network architectures, as well as with up-to-date concepts in the DL field. It is
important to highlight that the project is eminently applied, henceforth we foresee mostly the
implementation of a useful methodology rather than the advancement of the state-of-the-art.
Still, we expect publishing the methodology as an open-source software, or its usage as the
backend for a more advanced development.

2 Additional information

- **Difficulty of the project:** From moderate to very challenging.

- **What will you learn?** Natural language processing, convolutional neural networks,
  language models, image processing.

- **Requirements:** Good python level, knowledge of either Keras or PyTorch, experience
  with Git, machine learning fundamentals, creative thinking, some experience on the de-
  velopment of Dashboards and UIs.

- **Supervisors:** Luis Salamanca (luis.salamanca@sdsc.ethz.ch) and Fernando Pérez-Cruz
  (fernando.perezcruz@sdsc.ethz.ch).
Bibliography

References


