Towards Context-Aware Language Models for Historical OCR Post-Correction

December 4, 2018

1 Introduction

While the last decade brought substantial improvements in OCR, historic material is still challenging: special orthography, vocabulary, symbols, fonts, and layout, and sometimes poor visual quality hamper large-scale full-text digitisation. To account for these factors, on the language processing side of the problem, stochastic models need not only text samples but information which contexts they are representative of.

Owing to the vast diversity and complexity of language, context can span multiple dimensions at different scales: from global “domain” attributes like document type, genre, time, and origin to local “markup” information such as region class, font, form features, and language. Ideally, OCR should condition its approximation model on these meta-data as much as possible.

However, in practice such systems are already quite complex, meta-data schemes are very diverse, and historical ground truth corpora are still scarce [25]. So instead of extending existing OCR, it can be prudent to first build complementary tools for post-correction which feed from the OCR’s search space, but incorporate more linguistic knowledge, and maintain their own comprehensive models.

This project is part of the joint coordinated research effort OCR-D [1], which aims at developing OCR methods for historic material in preparation of automated mass-digitisation of all German-speaking prints from the 16–19th century. Our module is dedicated to post-correction. We shall first explore various techniques among the major modelling paradigms of weighted finite-state transducers (WFST) and artificial neural networks (ANN). Based on these experiments, a system will be implemented that can be executed within the OCR-D pipeline, or run standalone. We shall provide standard models which exploit all available data and meta-data for both historic and contemporary texts, as well as training tools.
2 Methodology

Many existing post-correction approaches employ WFST [27, 10, 18, 21, 3, 22], which allow for compact storage and processing of symbol sequences. Using prior knowledge on what mistakes OCR typically makes, how new words are formed grammatically, and which words are likely to appear next to each other, a post-correction system can be modelled as the composition of single transducers representing input character hypotheses, error model, lexicon, and word-level language model, each weighted stochastically. The result is a transducer containing all possible correction candidates alongside their probabilities.

In synthetic languages, a mere list of word forms does not suffice as word model. We will therefore extend the lexicon to a dynamic word model by using a morphology transducer.

One of the most common errors in OCR is word segmentation. Hence, post-correction should search across multiple words at once. Since the WFST approach is unable to handle long inputs efficiently, we will implement a sliding window technique, recombining window outputs afterwards.

Recent advances in deep-learning methods and their success in diverse applications have spurred researchers on proposing new post-correction approaches based on ANN [4, 12, 13]. Neural networks offer more efficient predictions, and are more powerful, while offering very good practical generalisation performance (the causes of which are slowly becoming understood [14, 6, 9, 20, 31, 2]).

Of particular attraction (especially in machine translation) is the encoder-decoder model for sequence-to-sequence transduction [11, 26] (often augmented with attention [5, 28], sometimes protected against exposure bias [29, 19, 16, 7]).

For estimation of the OCR error model, most WFST approaches use supervised learning, but unsupervised learning via Expectation Maximisation has been proven effective as well – both inductively [27] and transductively [22]. However, the data in all available contexts still must be split into a finite number of discrete, static domains, which arguably is an ill-defined, reductionist notion [24].

With ANN, since models can be pre-trained on auxiliary tasks and then transferred as initializer or fixed parameter sub-space [8, 15], it is easy to learn a single inductive model both on (scarce) labelled and (plenty) unlabelled data. Moreover, this model can be made context-aware by feeding all context variables as extra input and having the network learn an implicit lower-dimensional embedding, which dynamically combines with all components and thus makes them adaptive [17, 30].

Applied to the encoder-decoder post-correction model, this means we can perform supervised training by presenting OCR output to the encoder and ground truth text to the decoder, while also doing unsupervised training by presenting (both historic and contemporary)

- clean texts to both sides,
- noisy OCR texts to both sides, and
• clean text to the decoder, but randomly degraded text to the encoder [12].

Context variables could be input to both encoder and decoder, both hidden and output layers, both additively and multiplicatively.

3 Experiments

Early evaluation of our prototypes was done on GT4HistOCR’s dta19 sub-corpus [25]. Measuring CER and precision-recall [23], we find that

• character-level ANN outperform word-based WFST even when providing an oracle lexicon,

• ANN lean towards precision, whereas WFST allow some control over precision vs recall.

References


[25] Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. GT4HistOCR: Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin, August 2018.


