Correcting OCR errors for German in Fraktur font

Michel Généreux, Egon W. Stemle, Lionel Nicolas, Verena Lyding and Katalin Szabò

Introduction

The OPATCH project (Open Platform for Access to and Analysis of Textual Documents of Cultural Heritage) aims at creating an advanced on-line search infrastructure for general and linguistic research in text archives and corpora, including an historical newspaper archive.

For implementing this, OPATCH builds on computational linguistic (CL) methods for structural parsing, word class tagging and named entity recognition.

Dating between 1910 and 1920, the newspapers are typed in the blackletter Fraktur font and paper quality is derogated due to age.

Hence, in OPATCH we are starting from majorly error-prone OCR-ed text, in quantities that cannot realistically be corrected manually.

Corpora

Gold standard (GS): ten OCR-ed pages with their manually corrected versions

- 10,468 tokens and 3,621 types. Eight pages (8,324/2,487) are used as training data and two pages (2,144/1,134) for testing.
- More than one out of two tokens is misrecognized, among which almost half (48%) need a minimum of three edit operations for correction.

Reference corpus: 5M unigrams and 5M bigrams

- From the WEB and also http://www.gutenberg.org/

- Romane und Erzählungen (1910-20)

- The dictionary covers 91% of all words in the ten OCR-ed pages

Approach for OCR correction

- Probability models. Collate and tally all edit-operations (delete, insert and replace) needed to transform all unrecognized tokens from the training OCR-ed texts to its corrected form in the Gold Standard: e.g. n|u 98

- * we obtain two models: constrained (only edits seen in training) and unconstrained

- Candidate generation is achieved by finding the closest entry in the dictionary by applying the minimum number of edit-operations to an unrecognized OCR-ed token. The number of candidates is function of the maximum number of edit-operations allowed and the model used.

- * wundestc → wundesten: 'c' → 'e' and 'n' after the 'e'

- Selection of the most suitable candidate, given relative frequency and context. We maximize the function:

\[ \prod_{i} \text{prob}(\text{edit\_ops}) + \text{prob}(c) + \text{prob(left\_word\_c)} + \text{prob(c + right\_word)} \]

Experiment 1: Artificial errors

- To achieve this we extracted random trigrams from the GS and applied, in reverse, the edit error model.

- Errors were introduced up to two per target and contexts.

- At the end of this process, we have two context words and five candidates, including the target.

- En is the maximum edit-operations used to generate candidates.

- Table 1: Accuracy of OCR correction on 2,363 artificially created errors

<table>
<thead>
<tr>
<th>Five cand.</th>
<th>On the fly, open list of candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constr. model</td>
</tr>
<tr>
<td></td>
<td>E1</td>
</tr>
<tr>
<td>93%</td>
<td>86%</td>
</tr>
</tbody>
</table>

Experiment 2: Real errors

- Table 2: Accuracy of OCR correction on 233 real errors

<table>
<thead>
<tr>
<th>Constr. model</th>
<th>Unc. model</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>E2</td>
</tr>
<tr>
<td>GS</td>
<td>2153</td>
</tr>
<tr>
<td>[OCR]</td>
<td>2322</td>
</tr>
<tr>
<td>[GS ∩ OCR]</td>
<td>1185</td>
</tr>
<tr>
<td>[GS ∩ AC]</td>
<td>1268</td>
</tr>
<tr>
<td>Improvement</td>
<td>9%</td>
</tr>
</tbody>
</table>

Table 3: Error correction rate. | | size of

Results

- The approach we presented to correct OCR errors considered four features of two types: edit-distance and n-grams frequencies.

- Results showed that a simple scoring system can correct with very high accuracy OCR-ed texts under idealized conditions: no more than two edit operations and a wide-coverage dictionary.

- Obviously, these conditions do not always hold in practice, thus an observed accuracy drops to 10%. Wrong substitutions by the OCR process have also been neglected.

- Nevertheless, we can expect to improve our dictionary coverage so that very noisy OCR-ed texts (i.e. 48% error with distance of at least three to target) can be corrected with accuracies up to 20%.

- OCR-ed texts with less challenging error patterns can be corrected with accuracies up to 61% (distance two) and 86% (distance one).