Reducing OCR errors in Gothic-script documents

Furrer, L; Volk, M
Reducing OCR errors in Gothic-script documents

Abstract

In order to improve optical character recognition (OCR) quality in texts originally typeset in Gothic script, an automatic correction system can be built to be highly specialized for the given text. The approach includes external dictionary resources as well as information derived from the text itself.
Reducing OCR Errors in Gothic-Script Documents

Lenz Furrer and Martin Volk
Institute of Computational Linguistics, University of Zurich
Lenz.Furrer@uzh.ch, volk@cl.uzh.ch

July 14, 2011

Abstract

In order to improve optical character recognition (OCR) quality in texts originally typeset in Gothic script, an automatic correction system can be built to be highly specialized for the given text. The approach includes external dictionary resources as well as information derived from the text itself.

The resolutions by the Zurich Cantonal Government from 1887 to 1902 are archived as printed volumes in the State Archive of Zurich. The documents are typeset in Gothic script, also known as blackletter or Fraktur. As part of a collaborative project between the State Archive and the Institute of Computational Linguistics at the University of Zurich, these texts are being digitized for online publication. The aims of the project are automatic image-to-text conversion (OCR) of the approximately 11,000 pages, the segmentation of the bulk of text into separate resolutions, the annotation of metadata (such as the date or the archival signature) as well as improving the text quality by automatically correcting OCR errors. From an OCR perspective, the data are most challenging, as the texts contain not only Gothic type letters – which lead to a lower accuracy compared to antiqua texts – but also particular words, phrases and even whole paragraphs printed in antiqua font. Although we were lucky to have available an OCR engine capable of processing mixed Gothic and antiqua texts, the alternation of the two fonts still has an impairing effect on the text quality. Since the interspersed antiqua tokens can be very short (e.g., the abbreviation ‘Dr.’), their diverting character is sometimes not recognized. This leads to badly misrecognized words due to the quite different shapes of the typefaces; for example antiqua ‘Landrecht’ (Engl.: citizenship) is rendered as completely illegible ‘I>aier<ireclitt’, which is clearly the result of using the inappropriate recognition algorithm for Gothic script.
The main emphasis of the RRB-Fraktur project is on the post-correction of recognition errors. The evaluation of a limited number of text samples yielded a word accuracy of 96.96%, which means that one word out of 33 contains an error (for example, ‘Regieruug’ instead of correct ‘Regierung’, Engl.: government). We aim to significantly reduce the proportion of misrecognized word tokens by identifying them in the OCR output and determining the originally intended word with its correct spelling. The task resembles that of a spell-checker as found in modern text processing applications, with two major differences. First, the scale of the text data (with an estimated 11 million words more than 300,000 erroneous word forms are to be expected) does not allow for an interactive approach, asking a human user for feedback on every occurrence of a suspicious word form. Therefore we need an automatic system that can account for corrections with high reliability. Second, dating from the late 19th century, the texts show historic orthography, which differs in many ways from the spelling encountered in modern dictionaries (eg historic ‘Mittheilung’ versus modern ‘Mitteilung’, Engl.: message). This means that using modern dictionary resources directly, cannot lead to satisfactory results. Additionally, the governmental resolutions typically contain many toponymical references, which are not covered by general-vocabulary dictionaries. Regional peculiarities are also evident, for instance pronunciation variants or words not common elsewhere in the German speaking areas (eg ‘Hülfstrupp’ versus the standard German ‘Hilfstruppe’, Engl.: rescue team), and of course there is a great amount of genre-specific vocabulary, ie administrative and legal jargon. We are hence challenged to build a fully-automatic correction system with high precision that is aware of historic spelling and regional variants and contains geographical and governmental language.

The core task of the desired correction system with respect to its precision is the categorization routine that determines the correctness of every word. For example, ‘saumselig’ (Engl.: dilatory) is a correct word, whereas ‘Gefundheit’ is not (in fact, it is misrecognized for ‘Gesundheit’, Engl.: health). We use a combination of various resources for this task, such as a large dictionary system for German that covers morphological variation and compounding (such as ‘ging’, a past form of ‘gehen’, Engl.: to go, or ‘Niederdruckdampfsystem’, a compound of four segments meaning low-pressure steam system), a list of local toponyms, the recognition confidence values of the OCR engine and more. Every word is either judged as correct or erroneous according to the information we gather from the various resources. The historic and regional spelling deviations are modelled with a set of handwritten rules describing the regular differences. For example, with a rule stating that the sequence ‘th’ corresponds to ‘t’ in modern spelling, the standard form ‘Mitteilung’
can be derived from old ‘Mittheilung’. While the latter word is not found in the dictionary, the derived one is, which allows for the assumption that ‘Mittheilung’ is a correctly spelled word.

The set of all words recognized as correct words and their frequency can now serve as a lexicon for correcting erroneous word tokens. This corpus-derived lexicon is naturally highly genre-specific, which is desirable. On the other hand, rare words are likely to occur only in a misspelled version, in which case there will be no correction candidate in the lexicon. Due to the repetitive character of the text’s topic there is also a lot of repetition in the vocabulary across the corpus. This increases the chance that a word misrecognized in one place will have a correct occurrence in another. In an evaluation the approach has shown promising results. The text quality in terms of word accuracy could be improved from 96.96 % to 97.73 %, which is equivalent to a 25 % reduction in the rate of word misrecognition. As the system is far from complete, considerably improved results can be expected.

With this project we demonstrate that it is possible to build a highly specialized correction system for a specific text collection. We are using both general and specific resources, while the approach as a whole is widely generalizable. We see our work as an auspicious method for improving text quality of historic OCR-converted text.