Searching in text databases with non-standard orthography

Thomas Pilz, Universität Duisburg-Essen

In this paper we present research results of the recent project “Rule based search in text databases with non-standard orthography”. The interdisciplinary project is funded by the German Research Foundation.
Even though the digitization of text documents is a standard procedure nowadays, several problems remain to be solved. Low quality analogue copies demand preprocessing steps, involving binarization, component analysis, skew correction and de-warping (Mischke & Luther 2005). But even with most elaborate algorithms, recognition errors cannot be totally avoided if the sources lack a certain quality. Similarly, historical documents in black letter fonts cause recognition errors. Often, archives and retro digitization projects have to economize on costs of labor and therefore reduce or even omit manual correction. They maintain electronic documents with 40 and more percent of misinterpreted words. Texts authored well before the German standardization of orthography in 1901 may contain up to 60 additional percent of non-standard spellings (Kempken et al. 2006). Similar problems are documented for numerous other European languages as well, among those Dutch, English, French and Slovenian.
To deal with those uncertainties in spelling, a Java-based toolkit is being developed. In combination with the open-source search engine Apache Lucene (http://lucene.apache.org) it has already been successfully applied to the online Nietzsche-Archive (Biella et al. 2003). Currently, a database of evidences for different types of spelling variation is being built. Constantly growing it presently features 12.891 word pairs of variants and their related standard expressions from 106 different texts. These originate from all over the German-speaking area and range from 1293 to 1919. The spelling variants therefore cover diachronical language development, diatopical variation, differences in transcription and evidences of optical character recognition errors. Among the latter are variants from antiqua as well as black letter sources. At the moment these data are clustered by the two parameters *timeframe* and *location* and an additional *category* flag. *Timeframe* depicts four significant stages in the development of the German language: Late Middle High German (1250 - 1350), Elder Early New High German (1350 – 1450), Younger Early New High German (1450 – 1650) and New High German (1650 – 1900). *Location* is divided in the commonly used regions Upper German (south of Speyer line), Central German (south of Benrath line but north of Speyer line) and Low German (north of Benrath line) while *category* indicates OCR/Non-OCR errors. Using a trainable string edit distance on the base of the expectation-maximization algorithm and our database as a training set (Kempken et al. 2006), we were able to train modular distance measures.

Dagstuhl Seminar Proceedings 06491
Digital Historical Corpora- Architecture, Annotation, and Retrieval
http://drops.dagstuhl.de/opus/volltexte/2007/1053

In a recent Evaluation a distance measure trained on 3500 New High German evidences achieved a recall between 0.900 (first relevant result on position one), and 0.983 (first relevant result is within the five top positions). Every historical spelling can be found on the average position 1.98. The precision for 100% recall of all historical variants is 0,69.
As soon as more evidences for recognition errors are available it is planned to use an OCR-trained measure to further enhance the quality of a partial text recognition algorithm for documents in black letter font, that also was developed in our institute.
In addition to their application in search engines, our measures have shown to be of use in cross-language comparison as well (Archer et al, 2006). The assignment to document categorization looks promising but remains to be investigated in detail.
We propose the application of our techniques to synoptic view interfaces to keep the users unaware of the underlying retrieval. A corresponding interface module is under development.

References

Biella, D., Dyllong, E., Kaiser, H., Luther, W., and Mittmann, T. (2003). *"Edition électronique de la réception de Nietzsche des années 1865 à 1945"*. Proceedings ICHIM03 015C, Paris, 8-12 Sep 2003

Kempken, S., Luther, W., Pilz, T. (2006). *Comparison of distance measures for historical spelling variants*. Proceedings IFIP AI 2006, Santiago de Chile, 20-25 Aug 2006

Mischke, L., Luther, W. (2005). *Document Image De-Warping Based on Detection of Distorted Text Lines*. ICIAP 2005, LNCS 3617, pp. 1068-1075. 2005