Using an Alignment-based Lexicon for Canonicalization of Historical Text

This paper addresses issues in orthographic normalization of historical German text. We investigate the utility of a finite deterministic canonicalization lexicon (also known as a "witnessed dictionary") semi-automatically constructed from a corpus of historical and contemporary editions of the same texts by comparing its performance on a simulated information retrieval task to that of a robust generative finite-state canonicalization architecture or "hypothetical dictionary", as well as that of a hybrid method which uses a finite lexicon to augment a generative canonicalization architecture.