Project 4 Algorithm Implementation

Group 3

Outline

- Error Detection Using SVM
  - Feature Extraction
  - Model Tuning
- Error Correction with Binary N-grams
- Evaluation and Performance

Project Objective

- Optical Character Recognition (OCR) software is widely available in proprietary (AbbyReader) and open-source (Google Tesseract) form
  - Performance is often disappointing even with additional training
    - Experience with Columbia's HistoryLab
- Given a set of poorly OCR'ed documents and their ground truth pairs, can we improve the **detection** and **correction** of incorrect terms?
  - Detection using SVM (paper D3)
  - Correction using Binary N-grams (C1)

Error Detection

Recognizing Garbage in OCR Output on Historical Documents

Richard Wudtke
CIS – University of Munich
wudtke@cis.uni-muenchen.de

Christoph Ringlstetter
CIS – University of Munich
kristof@cis.uni-muenchen.de

Klaus U. Schulz
CIS – University of Munich
schulz@cis.uni-muenchen.de

Error Detection

- Feature Extraction based on series of rules
  - Length of token, ratio of vowels to consonants, …
- Unique Bigram Frequency feature
  - Attempt to capture the likelihood of two characters appearing together
  - ‘ed’ vs. ‘yy’

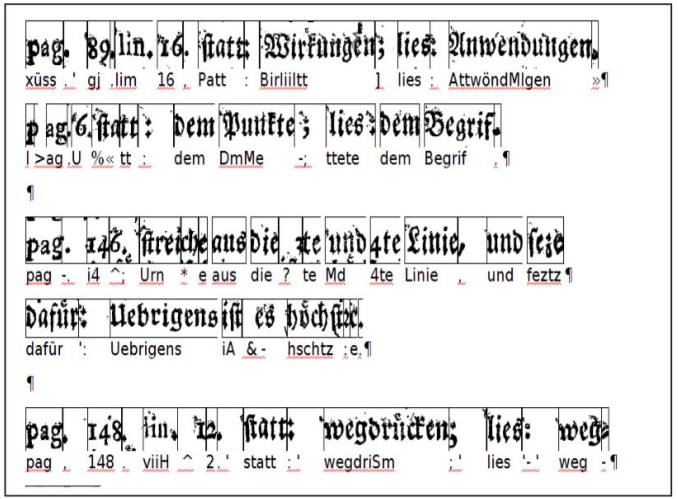

Figure 1: Images and OCR result aligned showing garbage and non-garbage tokens.

Error Detection

- Feature rules were developed with historical German texts in mind,
  - Higher likelihood of non-alphanumeric characters
- Some features related to non-alphanumeric characters were dropped at model fitting due to constant values

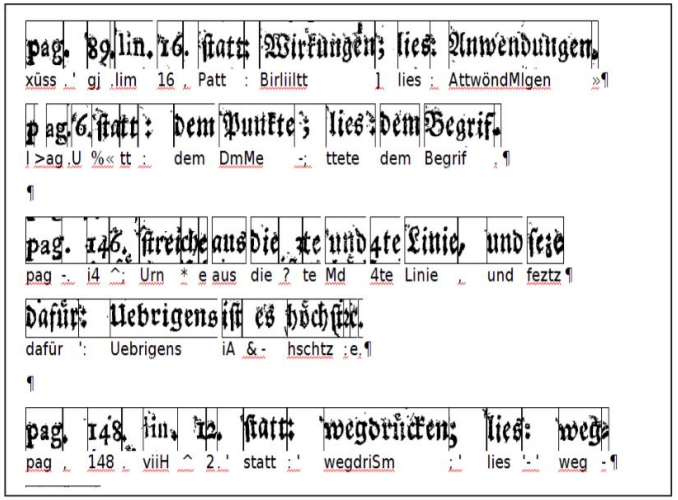

Figure 1: Images and OCR result aligned showing garbage and non-garbage tokens.

Error Detection

- Model fit using LIBSVM implementation in e1071 package
  - radial kernel
  - Training set of 75%, test set 25%
- Hyperparameter Tuning (gamma and cost)
  - Grid search implementation
    - Runtime considerations
  - Bayesian Optimization
    - Search over posterior distribution of functions
- **Ultimate detection rate of ~80%**

Error Correction

A Contextual Postprocessing System for Error Correction Using Binary *n*-Grams

EDWARD M. RISEMAN, MEMBER, IEEE, AND ALLEN R. HANSON, MEMBER, IEEE

Performance and Evaluation