Vortrag von Michael Strohmayer, BA Betreuer Dr. Klaus Schulz
Thema: Machine-Learning basierte automatische OCR-Korrektur

Als Motivation für die Arbeit waren folgende Punkte ausschlaggebend: Zum einen erkennen OCR Systeme manche Wörter nicht zuverlässig. Die Systeme liefern Korrekturvorschläge von denen der korrekte ausgewählt werden muss. Zum anderen unterliegen Schriftbilder, Grammatik und Schreibweisen einem stetigen Wandel.
Ziel ist es, ein „machine-learning" System zu trainieren, welches die automatische Nachkorrektur der eingelesenen OCR-Dokumente so korrekt wie möglich ausführt.
Um dieses Ziel zu erreichen, werden als erstes Dokumente eingelesen. Hier werden die verfügbaren Ground-Truth Dokumente „Paradiesgärtlein" und „Curiöser Botanicus" verwendet. Die Ground-Truth Dokumente werden hier ähnlich einem Goldstandard benutzt. Der RIDGES Korpus (33 Kräuterkundetexte aus den Jahren 1484 - 1914) wird für das Training verwendet. Im nächste Schritt werden gegebene Featurewerte extrahiert. Zu diesen Werten gehören die Levenshtein-Disztanz, die absolute Häufigkeit des OCR Token und eine Gewichtung welche eine Wahrscheinlichkeit für einen der zu ersetzenden Buchstaben angibt, sodass wieder ein sinnvolles Wort entsteht. Weitere Features die verwendet werden können sind die Längendifferenz zweier Token, der Konfidenzwert des folgenden Korrekturvorschlags und Frequenzlisten von Tokens.
Zum Training von „Machine-Learning" Klassifikatoren wird „Scikit-learn" (eine große Bibliothek von „Machine-Learning" und „Data Mining Tools") und „Libsvm" (nutzt Stützvektor-Maschinen zur Klassifikation) verwendet. „Scikit-learn" wird auch für den Naive Bayes Klassifikator benutzt.
Probleme gab es zu Beginn in der Datenverarbeitung. Das Programm schnitt die Konfidenzwerte in der Ausgabe ab und lieferte deshalb falsche Trainingsdaten. Das Problem wurde aber im Laufe der Arbeit behoben.
Die Evaluation ergab, dass die Berechnung von Naive Bayes sehr schnell ist, „libsvm" jedoch hochwertigere Ergebnisse bietet. Kreuzvalidierung erzielte insgesamt bessere Ergebnisse.
Als Bewertungsmaße wurden Precision, Recall, F1 und Accuracy verwendet. Nutzt man nur die Basisfeatures zur Korrektur, sind Precision und F1 sehr niedrig. Mithilfe der Zusatzfeatures können Precision und F1 jedoch deutlich verbessert werden und auch die Accuracy wird noch leicht verbessert. Recall bleibt sowohl mit Basis- und Zusatzfeatures konstant.
Häufige Fehlklassifikationen traten auf, wenn in den Ground-Truth Daten kein Korrekturvorschlag vorhanden war oder der Profiler einen falschen Konfidenzwert geliefert hat.
Insgesamt hat sich die automatische OCR Erkennung als sinnvoll erwiesen, da gute Ergebnisse erzielt wurden. Um das System zu erweitern, könnte man Klassifikatoren kombinieren und weitere Features hinzufügen.
Zum Schluss des Vortrags wird noch auf vertiefende Literatur hingewiesen.

Verfasser: Thomas Ebert