CIS, LMU München
Fach: Kolloquium Computerlinguistisches Arbeiten SS2017
Studentin: Faridis Alberteris Azar

Protokoll zur Sitzung am 19.06.2017

Thema: Machine-Learning basierte automatische OCR-Korrektur

Student: Michael Strohmayer
Betreuer: Dr. Klaus Schulz

Die Tatsache, dass sich die Schriftbilder, Grammatik und Schreibweisen verändern und daher erkennen OCR-Systeme manche Wörter nicht zuverlässig. Sie liefern eine Liste an Korrekturvorschlägen von denen der korrekte ausgewählt werden muss. Der Student erstellt in dieser Arbeit eine Software zur automatischen Nachkorrektur der eingelesenen OCR-Dokumente.

Die Dokumente werden eingelesen und die gegebenen Feature-Werte extrahiert. Im Nachhinein werden neue Features hinzufügt und die verwendeten Machine-Learning Klassifikatoren trainiert.

Michael verwendet Dokumente wie die verfügbare Grund-Truth Dokumente „Paradiesgärtlein“ und „Curiöser Botanicus“. Er benutzt außerdem das RIDGES Korpus, das 33 Kräuterkundetexte aus der Zeit zwischen 1484 und 1914 enthält und von CIS LMU in Zusammenarbeit mit Humboldt Universität in Berlin erstellt wurde.

Die neuen Features, die Strohmayer erzeugt und hinzugefügt hat, sind:

- Längendifferenz
- Konfidenzwert des folgenden Korrekturvorschlags (Leider nicht so viel gebrach wi er erhofft hat.)
- Frequenzlisten

Er benutzt für die Klassifikation Scikit-learn und Libsvm. Scikit-learn hat eine große Bibliothek an Machine-Learning und Data Mining Tools und verwendet den Gauß Naive Bayes Klassifikator, von dem die wir bereits in den Statistikvorlesungen gehört haben. Libsvm verwendet Support Vector Machine zur Klassifizierung von Daten.

Der Student hat einige Probleme während der Arbeit konfrontiert, wie zum Beispiel Perfomance Probleme in der Datenverarbeitung, was er mit der Verwendung von einem Dictionary gelöst hat. Außerdem wurde eins von seinen neuen Features, Konfidenzwert, in der Ausgabe abgeschnitten, was zu falschen Trainingswerten geführt hat. Das war deswegen einer seiner Klassifikationsfehlern.

Für die Evaluation hat der Referent beide Klassifikatoren vergliechen. Der Naive Bayes Klassifikator war sehr sehr schnell im Vergleich zu dem anderen. Aber der andere Klassifikator hat bessere Ergebnisse geliefert.

Der Student schlagt für zukünftige Arbeiten eine Kombination von beiden Klassifikatoren und die Erzeugung von weiteren Features vor.