Zusammenfassung zum Vortrag über die Bachelorarbeit zum Thema „Machine Learning basierte automatische OCR-Korrektur“ von Michael Strohmayer, zusammengefasst von Korbinian Schmidhuber

Die Motivation für diese Arbeit lag darin, dass OCR-Systeme manche Wörter nicht zuverlässig erkennen, was zum Beispiel aufgrund von Veränderungen in Schriftbildern, der Grammatik, oder der Schreibweisen in einer Sprache passieren kann. OCR-Systeme liefern in solchen Fällen eine Liste von Korrekturvorschlägen, von denen der Benutzer dann den korrekten Vorschlag auswählen muss.

Ziel der Arbeit war es deshalb eine Software zu erstellen, die mithilfe eines Klassifikators eine automatisierte Nachkorrektur der eingegebenen OCR-Dokumente durchführt.

Die Vorgehensweise war hierbei zunächst, die Dokumente einzulesen, Featurewerte zu extrahieren, eigene Featurewerte hinzuzufügen, das Training des Klassifikators und anschließend eine Evaluation dessen.

Herr Schulz erklärt an dieser Stelle kurz die Bedeutung eines Profilers: Ein Profiler nimmt die Eingabedateien und erstellt Interpretationen der einzelnen Begriffe, die er z.B. dadurch erhält, in dem er überprüft, wie ein Wort in diesem Textzusammenhang heißen müsste, oder die Verwendung paralleler moderner Wörter, usw. Anschließend liefert er eine Liste gewichteter Interpretationen.

Diese gewichteten Interpretationen sollen als Eingabe für die entwickelte Software benutzt werden. Als Dateien wurden Dokumente aus dem RIDGES Korpus verwendet, welches aus Kräuterkundetexten in einem Zeitraum von 1484 bis 1914 erstellt wurden. Dieses Korpus wurde am CIS, in Zusammenarbeit mit der Humboldt Universität Berlin erstellt.

Nachdem das Dateiformat der Ausgabe des Profilers und deren Featurewerte vorgestellt wurden, die extrahiert wurden, wird aufgezeigt, welche Featurewerte für den Klassifikator zusätzlich noch hinzugezogen werden sollen: Die Längendifferenz zwischen dem OCR-Token und dem Korrekturvorschlag, der Konfidenzwert des folgenden Korrekturvorschlages (welcher allerding weniger Mehrwert für den Klassifikator eingebracht hat, als erwartet) und die Frequenz, die allerdings später nicht mehr gebraucht wurde.

Als Tools zur Klassifikation wurden einerseits scikit-learn verwendet, ein Tool das den Gauß-Naive Bayes Algorithmus zur Klassifikation verwendet, sowie libsvm, ein Toll welches eine Support Vector Machine zur Klassifikation benutzt.

Ein Problem, welches im Rahmen der Arbeit aufkam, war eine sehr schlechte Performanz, die allerdings mit einer geeigneteren Datenstruktur deutlich verbessert werden konnte.

Bei der Evaluation ergab sich, dass eine cross-validation zu deutlich besseren Ergebnissen führt. Zudem wurden die Ergebnisse der Tools zur Klassifikation miteinander verglichen. Hierbei lieferte scikit-learn um einiges schnellere Ergebnisse, die allerdings verglichen mit libsvm deutlich schlechter ausfielen. Die Bewertungsgrößen, die für die Evaluation benutzt wurden waren: Precision, Recall, Accuracy und F1 Score.

Fehler kamen vor allem dann zustande, wenn der Profiler falsche Konfidenzwerte liefert.
Als Ausblick und Fazit führte Michael auf, dass eine automatische Nachkorrektur sehr sinnvoll ist und gute Ergebnisse erzielt. Diese könnten durch Hinzunahme weiterer Features noch weiter verbessert werden, ebenso durch die Kombination beider Klassifikatoren