Machine-Learning basierte automatische OCR-Korrektur

Der zweite Vortrag des Tages wurde von Michael Strohmayer gehalten. Sein Betreuer ist Dr. Klaus Schulz. Das Ziel der Arbeit ist die automatisierte Korrektur von OCR-Dokumenten. Die OCR-Erkennung ist manchmal fehlerhaft und unzuverlässig, weshalb eine solche Korrektur Sinn macht. Dabei liest er die entsprechenden Dokumente ein und trainiert, mit Hilfe von einigen „Features“, ein Machine Learning System. Er hat verschiedene Korpora verwendet, wie zum Beispiel den RIDGES Korpus oder die Dokumente „Paradiesgärtlein“ und „Curiöser Botanikus“. Wenn also diese Dokumente eingelesen wurden, werden die Featurewerte extrahiert und ein paar weitere Features dazu gefügt. Beispiele für diese Features sind die Längendifferenz, Frequenzlisten und der Konfidenzwert des folgenden Korrekturvorschlags. Mit diesen Features lässt sich jetzt der Machine-Learning Klassifikator trainieren. Es gibt zwei dieser Klassifikatoren. Einmal den Scikit-learn, welcher eine Bibliothek an Machine-Learning und Data-Mining Tools ist. In diesem Rahmen wurde ein „Gauß Naive Bayes Klassifikator“ verwendet. Der andere Klassifikator ist der Libsvm, dieser verwendet Support Vector Machines, um die Daten Klassifizieren. Während der Arbeit sind einige Probleme aufgetreten. Eines davon ist ein Performanz Problem. Ein weiteres Problem waren fehlerhafte Konfidenzwerte, welche zu falschen Trainingswerten geführt haben. Um bessere Ergebnisse zu erzielen wurde eine Kreuzevaluation verwendet. Außerdem erwies sich der Naive Bayes Ansatz als sehr schnell, jedoch schaffte der libsvm Klassifikator bessere Ergebnisse. Weitere Schritte wären zum einen die Kombination der beiden Klassifikatoren gewesen und zum andere hätte man noch weitere Features hinzunehmen können.

Musik und Ludwig Wittgenstein: Semantische Suche in seinem Nachlass

Das erste Referat wurde von Ines Röhrer gehalten. Ihr Betreuer ist Max Hadersbeck. Diese Arbeit befasste sich mit der Suchmaschine WiTTFind und dem Nachlass von Ludwig Wittgenstein. Nur ein Teil des Nachlasses ist der Öffentlichkeit zugängig. WiTTFind wurde extra für diesen Teil des Nachlasses erstellt und beinhaltet sowohl eine semantische Suche, als auch eine regelbasierte Suche. Der Nachlass selbst besteht aus vielen verschiedenen Bemerkungen, die zum Teil Manuskripte und zu Teil Typoskripte sind. Das Ziel der Arbeit ist die Erweiterung von WiTTFind, in Richtung der semantischen Suche. Im Vordergrund steht die Musik, da diese für Wittgenstein sehr wichtig war. Ein Teil der Arbeit war also ein Musiklexikon zu erstellen. Dabei hatte sie Zugriff auf eine Hausarbeit, die viele wichtige Begriffe zu diesem Thema beinhaltete. Zur besseren Unterscheidung wurden die Begriffe in verschiedene Kategorien unterteilt, zum Beispiel „Instrumente“ und „Komponisten“. Zusätzlich hat sie Frequenz-Berechnungen durchgeführt. Im Laufe der Arbeit geriet der Kontext, in dem die Begriffe standen in den Fokus der Arbeit. Um den Kontext zu erfassen hatte sie sich für zwei Verfahrensweisen entschieden. Zum einen verwendete sie einen Ringbuffer. Für jede Kontextvariante wurde ein eigener Ringbuffer erstellt. Jedoch war der Zeitliche Aufwand, bei dieser Methode sehr hoch. Zum anderen hat sie mit Listenoperationen gearbeitet. Diese Variante wurde dem Ringbuffer am Ende vorgezogen. Ein weiterer Teil der Arbeit, war die Darstellung von Relationen zwischen den Musikwörtern. Dies wird durch eine Ontologie erreicht. Eine bereits vorhandene Ontologie zu diesem Thema ist „The Music Ontology“. Mit Hilfe dieser Ontologie

hat sie einen kleinen Prototyp erstellt. Dieser kann in Zukunft ausgebaut werden.