Joseph Birkner
Mat.Nr. 11393024

Protokolle zum Colloquium (19.06.2017)

Studentin: Röder, Betreuer: Maximilian Hadersbeck

„Musik und Ludwig Wittgenstein – Semantische Suche im Nachlass“

Im Rahmen des Projektes „Wittgenstein in Co-Text“ widmet sich das Centrum für Informations- und Sprachverarbeitung (CIS) gemeinsam mit der Uni Bergen der Transkription, Analyse und Veröffentlichung des Nachlasses des Philosophen Ludwig Josef Johann Wittgensteins (1889-1951).

Hierbei entwickelt das CIS die Software WittFind, mit der annotierte Datensätze, die in XML-Form von der Uni Bergen aus den Nachlässen eingelesen werden, durchsucht werden können. Ein wichtiger Anwendungsfall von WittFind, der leider bisher nicht umgesetzt ist, ist die Auffindung von Erwähnungen von Musikbegriffen in den Transkriptionen.

Die Auffindung der Erwähnungen von Musikbegriffen ist hierbei Teil der s.g. „Semantischen Suche“ - d.h. Suchen nach bestimmten Themen oder Konzepten anstatt von syntaktischen/lexikalischen Merkmalen. Ein Beispiel für ein solches Semantisches Modul das bereits in die Suche integriert ist, ist das s.g. „Farbmodul“ - hier kann nach nach Erwähnungen bestimmter Farben oder auch Farbkategorien wie „Transparenz“ oder „Zwischenfarben“ gesucht werden.

Ein erster Schritt in der Entwicklung des Moduls für Musikbegriffe war die Feststellung der Begriffe, die in der Auswahl verfügbar sein sollen. Hierfür wurde die externe Hilfe eines Musikwissenschaftler aus der LMU in Anspruch genommen, der folgende Bagriffskategorien determinierte: Komponisten, Gattungen, Intervalle, Kompositionsbezüge, Sonstige. Über diese Kategorien sind 98 Begriffe verteilt, die von dem Modul erfasst werden.

Probleme ergaben sich in verschiedener Hinsicht: Das für die Auffindung der Begriffe herangezogene Vollformenlexikon zum Beispiel erwies sich für die spezielle musikalische Terminologie als unzuverlässig, statt dessen musste eine handgefertigte Endungsliste herangezogen werden. Der Aufbau einer Ontologie (Graph aus Objekten und Beziehungen) über den aufgefundenen musikalischen Begriffen erwies sich als den Rahmen der Arbeit sprengend.

Student: Michael Strohmaier, Betreuer: Klaus Schulz

„Machine-Learning basierte automatische OCR-Korrektur“

Optical Character Recognition (OCR) spielt eine wichtige Rolle bei der Digitalisierung von Dokumenten, sowohl typografischer als auch handgeschriebener Natur. Die hierbei eingesetzten Systeme bezeichnet man auch als „Profiler“, da sie Ursprünglich eine eher unterstützende Rolle bei der Transkription einnahmen. Mit wachsender Rechenleistung und Verbesserungen in Computer-Vision Algorithmen ist es jedoch nun denkbar, vollauttomatische OCR-Systeme einzusetzen.

Ein großes Problem hierbei ist der Umgang mit Ambiguitäten im Schriftbild, bei denen der Profiler für keine Alternative eine hinreichend große Wahrscheinlichkeit berechnen kann. An solchen Stellen liefert der Profiler statt dessen eine Liste von alternativen Vorschlagswörtern mit ihrer jeweiligen Wahrscheinlichkeit.

Um diese Ambiguität aufzulösen hat Michael verschiedene Classifier trainiert, welche die Alternativen ranken und und so die Ambiguität auflösen sollen. Als Input-Features erwähnte er die s.g. Basis-Features (Profiler-Konfideny, Levenshtein-Distanz, Worthäufigkeit) und die Extra-Features Konfindenz der nachfolgenden Alternative und Längenunterschied der Alternative. Trainiert wurde ein Naive-Bayes Classifier aus Scikit sowie eine Supportvektormaschine aus LibSVM. LibSVM erwies sich als genauer, jedoch langsamer.

Das beste Ergebnis einer F1-Score von 90% erzielte Michael mit dem Featureset Basis+NächsteVorschlagswortKonfidenz+Längenunterschied auf dem RIDGES-korpus.