ZUSAMMENFASSUNGEN COLLOQUIUM, REPETITORIUM

Pascal Guldener

19.6.2017

1 Colloquium

1.1 Musik und Wittgenstein - Semantische Suche Im Wittfind

DIe Bachelorarbeot von Ines Röhrer betreut von Dr. Hadersbeck, hatte zum Ziel das vom CIS entwickelten und betreuten Tool, Wittfind , eine Semantische Kategorie zur Suche im Nachlass von Ludwig Wittgenstein hinzuzufügen. Die Motivation ist die Annahme, dass Musik im Leben und Werk des Sprachphilsophen eine einschlägige Rolle gespielt habe und such das in seinem geistigen Nachlass wiederspiegele. Dies motivierte die Autorin zur Zusammenarbeit mit einem Musikwissenschaftler zur Identifizierung der relevantesten Begriffe aus der Domäne Musik im Nachlass von Wittgenstein. Zunächst erleuterte Sie die Verwendung semantischer Kategorien im Wittfind-tool anhand von der Kategorie "Farbe" welche eine Klasse von farbbeischreibenden Begriffen bildet. Als Ihre Aufgabe beschrieb Sie nun, diese Klasse analog für musikalische Begriffe zu definieren und eventuell auch eine Ontologie zu entwickeln die es erlaubt Relationen zwischen diesen Begriffen zu finden.

Die Basis ihrer Arbeit bildete alle Vollformenextraktion aller relevanten Begriffe. In Ermangelung geeigneter Tools hatte sie sich daür ein eigenes Vollformenlexikon geschrieben welches aus Endungslisten und denzugehörigen Lemmata der Kategorie bestand. Wärend für Sie die Nachteile und Grenzen diese Verfahrens auf der Hand liegen, stellte dies eine verläufig akzeptable Grundlage für Ihre Arbeit dar. Als weitere große Herausforderung stellte sich die Disambiguierung heraus. Frau Röhrer erleuterte dies anhand der verschiedenen Verwendungen des Buchstabens "C". Dieser kann als Variable in einer einem mathematischen Kontext genauso auftauchen wie bei der Bennennung des Tons in einem muskialischen Kontext. Um die Kontextsensitivität bei der Desambiguierung solcher Begriffe zu auszunutzen implementierte sie einen Ringpuffer in welchem auf semantisch verwandte Begriffe in einen gewissen Bastand geprüft werden kann.

Der grossteil des praktischen Teils bezog sich auf die technischen Herausforderungen, insbesondere die GUI des Wittfind um die notwendigen JavaScript-Komponenten zu ereitern welche für die praktische Verwendung nötig sind. Dabei kammen auch Unzulänglichkeiten in der bevorzugten Darstellung (linear in Abhängikeit zu Ihrer Häufigkeit) diser "Wortfelder", bzw. Klassen, als Wortcloud zur Sprache. Viele Wörter kamen so selten vor, dass sie kaum mehr wahrnehmbar dargestellt wurden während andere den Großteil des verfügbaren Platzes einnahmen.

Aus Zeitmangel kam es allerdings nicht mehr zur Ausarbeitung einer Ontologie und Modellierung der Ralationen von Begrifflichkeiten zueinander.

1.2 Machine-Learning basierte automatische OCR-Korrektur

Die Bachelorarbeit von Michael Strohmeier, betreut von Prof. Schulz, behandelt die Frage ob sich das Ergebnis einer OCR Korrektur durch mit Hilfe einer Klassifizierung durch ein Machine Learning System verbessern lässt. Als features verwendete er dabei die Ausgabe des Profilers des von ihm verwendet OCR-Systems. OCR-Systeme werden vor mannigfaltige Probleme gestellt, dazu gehören nicht nur die Erkennung von Handschriften oder Robustheit gegegenüber verschiedenen

Fonts sondern auch die Veränderung der korrekten Schreibweise eines Worts über die Zeit.
Diese Unwägbarkeiten machen es sehr schwierig ein System zu bauen welches keiner Korrektur durch Menschen bedarf. Die Ausgabe des verwendeten OCR-Systems bestand aus einer Liste von Korrekturvorschlägen mit Konfidenzwert und weiteren Informationen, wie dem Levenstheinabstand. Während diese vom OCR-System als interaktive Liste zur manuellen Auswahl zur Verfügung gestellt werden, wurden nun ein Teil dieser Informationen als Features für den Klasssifizierer verwendet. Weitere hinzugefügte Features waren die Längendifferenz, der Konfidenzwert des zweitgerankten Korrekturvorschlags und die Häufigkeit des betreffenden Wortes. Als Klassifikatoren wurden ein Naive-Bayes Klassificator aus der Scit-Learn Library sowie eine durch die Libsvm bereitgestellte Support Vector Machine verwendet. Das Goldstandard-Datenset bestand aus den Korpora "Pardiesgärtlein", "Curioser Botanicus" und dem aus 33 Kräuterkundetexten bestehenden RIDGES Korpus, welche vom CIS in Kooperation mit der Humboldt Universität Berlin erstellt wurden. Herr Strohmeier berichtete von anfangs eher schlechten Resultaten, die er allerdings auf eine Verkürzung des Konfidenzwerte in der Ausgabe des OCR-Systems zurückführen konnte. Einerseits konnte dies durch Kreuzevaluierung, also der Einführung eines Developementdatensets, ausgleichen, andererseits hat sich dadurch auch die starke Abhängigkeit vom Feature des vorberechenten Konfidenzwerts gezeigt. Auf Nachfrage an Herrn Schulz ob man daher nicht das OCR-System selbst bereits als (Multiclass) Klassifikator anzusehen sei, äusserte dieser sich skeptisch und verwies auf die geringe Zuverlässgkeit, welche ein Beibehalten des zu korregierenden Wortes oft als die besssere Wahl erscheinen lässt.
Beim Vergleich der beiden Klassifikatoren zeigten sich sehr grosse Unterschiede in der Performance was die Laufzeiten bei Training und Anwendung angehen. Während die SVM 12s für das Training und 3.26s für eine Binär-Klassifizierung benötige brauchte der Naive-Bayes Klassifikator dafür 0.32s und 0.009s. Da die SVM bessere Ergebnisse lieferte wurde Sie für die Evaluation herangezogen. Vergleiche über Basisfeatures und extrahierten Features zeigten einen deutlichen Performancegewinn mit einem finalen F-Score von über 80%. Herr Strohmeiers Fehleranalyse ergab, dass die Abdeckung der zu klassifizierenden Wörter durch den Goldstandard sowie die Qualität des Konfidenzwerts entscheidende Kriterien für die Performance der Klassifikatoren waren.

2 Repetitorium

2.1 Versionierung mit Git

Herr Roth ging auf die Enstehungsgeschichte von Git, insbesondere die Rolle von Git bei der Entwicklung des Linux-Kernels, Unterschiede zu älteren Versionierungsystemen und praktische Verwendung ein. Insbesondere illustrierte er die verschiedene Zustände die eine Datei in Git (add,stage,commit) anehmen kann und erklärte die fundamentalen Konzepte von Branching und Merging. Ausserdem empfahl er publickey authentication und erleuterte grob deren Einrichtung. Zuletzt führte er in grundlegende git Befehle wie git rm , git diff und die Funktionalität von .gitignore ein. Auch die Verwendung von diffs zum Vergleich alter mit neueren Versionen und die Commit-History wurden knapp erklärt.