Protokolle zur Sitzung am 19.06.2017

"Musik und Wittgenstein: Semantische Suche in seinem Nachlass"

Ines Röhrer, Betreuer: Dr. Maximilian Hadersbeck

Ines stellt zunächst das CIS LMU Tool "WITTfind" vor, das schon in anderen Arbeiten von Bedeutung war. Dabei handelt es sich um eine Art Suchmaschine, um den Nachlass von Wittgenstein entweder regelbasiert oder semantisch nach Informationen zu durchsuchen. Als Beispiel für eine semantische Suche, wird die Kategorie "Farbe" genannt. Wählt man diese, erhält man eine Frequenzliste der Vorkommen von Begriffen, die der Farb-Kategorie zugeordnet werden können. Die Ergebnisse der Suche werden mithilfe einer Word Cloud graphisch dargestellt. Diese Funktion hat Ines sich zum Vorbild genommen, äquivalent dazu eine Musik-Kategorie zugänglich zu machen. Für die Arbeit werden die öffentlich zugänglichen Dokumente aus dem Nachlass verwendet. Leider stellen diese Dokumente nur einen kleinen Teil, des gesamten Nachlasses dar. Somit werden viele Begriffe, die der musikalischen Kategorie zugehörig sein könnten, außer Acht gelassen. Zusätzlich zur Implementierung einer neuen, musikalischen Kategorie für WITTfind, wird versucht, die Relationen zwischen den einzelnen musikalischen Begriffen zu beschreiben. Dazu werden Ontologien der jeweiligen Begriffe untersucht. Ein lokaler Webserver dient hierbei zu Testzwecken für die Implementierung verschiedener Algorithmen, ohne dass der eigentliche WITTfind Server involviert wird. Das fertige Modul wurde anschließend in die HTML-Dateien von WITTfind integriert. Eine zur Verfügung gestellte Hausarbeit beschäftigt sich bereits mit der Einordnung von musikalischen Begriffen. Nun galt es, die Ergebnisse und Strukturen dieser Hausarbeit so zu überarbeiten, dass sie mit dem Ziel der Arbeit konform sind. Ines nennt Probleme wie zu niedrige Frequenz von Wörtern oder Ambiguität. Dazu wird ein weiteres Beispiel hinzugezogen: der Begriff "c" könnte die semantisch relevante Note darstellen, allerdings auch eine mathematische Variable beschreiben. Dr. Schulz wirft hier ein, dass Named Entity Kategorien bei der Erstellung von Ontologien durchaus helfen können. Die Musik-Kategorien, die aus dieser Arbeit entstanden sind, lauten wie folgt: Komponist, Instrument, Gattung, Intervalle, Bezug zu Komposition, Sonstige Begriffe. Für einen Laien kann eine eindeutige Zuordnung vielleicht schwierig erscheinen, doch diese Zuordnung ist unter Musikwissenschaftlern weitestgehend einheitlich. Dies beweist auch wie interdisziplinär eine solche Arbeit sein kann. Es wird eine beispielhafte Wordcloud präsentiert, die die Frequenz musikalischer Begriffe ausgeben soll. Einer der Zuhörer fragt sich, ob solch eine Darstellung ob der hohen Differenzen zwischen den Frequenzen Sinn macht und ob man das Darstellungsproblem (sehr groß dargestellte häufige Wörter und verschwindend kleine Darstellung für seltene Begriffe) nicht hätte logarithmisch lösen können. Ines erklärt, dass die graphische Darstellung nicht der Fokus ihrer Arbeit war. Stattdessen geht sie näher auf die Erstellung der Frequenzlisten ein, die dafür benötigt wurden. Hier wurde zunächst mit einem eigens erstellten, primitiven Endungslexikon gearbeitet, um mögliche Vollformen zu extrahieren (z.B. Lemma Brahm für die Form Brahm'sche). In einer vorangegangen Präsentation von Faridis Alberteris wurde ein Vollformenlexikon vorgestellt, das auch in dieser Arbeit zu einem späteren Zeitpunkt zum

Einsatz kam. Die Frequenz wird berechnet, in dem ein Begriff als Key eines Dictionaries angelegt wird mit der Frequenz und der Ursprungsdatei als Value. Die Probleme, die dabei entstanden sind, waren mitunter fehlende Leerzeichen, die in den Dokumenten auftreten. Der erste Gedanke, eine In-Text-Suche zu betreiben, wurden jedoch fallen gelassen, da nun viel irrelevantes mitaufgegriffen wird (musikalische Begriffe wie "Bach" können sich innerhalb anderer, nicht relevanter Begriffe wiederzufinden sein). Dieses Problem wurde jedoch durch bessere, saubere Korpora behoben. Eine andere Problematik war das sog. multiple Satzvorkommen. Hier war die Frage eher, ob Begriffe, die in identischen Sätzen vorkommen nun ignoriert werden sollten, oder zur Frequenz hinzugezählt werden. Für diese Arbeit wurden die multiplen Vorkommen akzeptiert mit der Begründung, dass diese wohl sehr wichtig zu schein scheinen und nicht redundant. Anschließend wird die Vorgehensweise bei der Extraktion von Kontexten zwischen den musikalischen Begriffen näher beschrieben. Zunächst ist der Rahmen, der für einen möglichen Kontext gewählt wurde auf 50 Wörter gesetzt, der sich aber im Laufe der Zeit auf 5 beschränkt. Ein Ringbuffer extrahiert den Kontext, sobald der Lesepointer einen relevanten Treffer erzielt. Listoperationen dienen dann zur Kontrolle von Relevanz des Kontexts. Die Erstellung einer Ontologie stellt sich als sehr komplex und zeitaufwendig heraus. Dennoch ist es Ines gelungen, einen Prototypen für die Kategorie "Komponist" mithilfe einer bereits vorhandenen (The Music Ontology) zu erstellen. Die Arbeit war insgesamt sehr interdisziplinär aufgebaut und auch die Ansicht Wittgensteins wurde miteinbezogen. Die Ergebnisse waren erfolgreich, wenn auch nicht perfekt.

"Machine-Learning basierte automatische OCR-Korrektur"

Michael Strohmeyer, Betreuer: Dr. Klaus Schulz

Michael beginnt seinen Vortrag mit der Motivation für seine Arbeit: Schrift und Grammatik ändern sich im Laufe der Zeit und das führt dazu, dass Texterkennungssysteme (OCR) Wörter nicht zuverlässig genug erkennen. Sobald Zweifel bei der Erkennung entstehen, wird eine Liste an möglichen Korrekturvorschlägen ausgegeben. Diese Vorschläge werden beispielsweise anhand von Levenshtein-Abständen generiert. Nun soll ein Machine Learning Klassifikator trainiert werden, um automatisch aus dieser Liste, den richtigen Treffer auszuwählen. Um so ein System auf die Beine zu stellen, werden Dokumente eingelesen, Featurewerte des Profilers extrahiert, neue Feature hinzugefügt und anschließend 2 Klassifikatoren trainiert. Dr. Schulz gibt eine kurze Beschreibung eines Profilers: ein Profiler gibt gewichtete Interpretationen aus und ermöglicht so mehr Korrekturvorschläge für ein vollautomatisches System. Zunächst wird der Input des Systems behandelt. Dabei handelt es sich um zwei Grund-Truth-Dokumente aus der Kräuterkunde: Paradiesgärtlein und Curiöser Botanicus. Der RIDGES-Korpus enthält 33 Kräuterkundetexte aus den Jahren 1484-1914 und wurde von der LMU und der Humboldt Universität Berlin erstellt. Dieser wurde händisch nachkorrigiert. Der Datei-Output beschreibt die Wahrscheinlichkeit, wie sicher das System sich der richtigen Antwort ist, die Levenshtein-Distanz, die Zeichen, die ersetzt werden

sollen, und das korrekte Token. Das System wurde dann um drei zusätzliche Features erweitert. Das erste soll die Längendifferenz zwischen dem Token und dem Korrekturvorschlag ausgeben. Das zweite soll den Konfidenzwert des folgenden Korrekturvorschlags in Betracht ziehen und das letzte ist eine Frequenzliste, die jedoch später von der des Profilers ersetzt wurde. Für diese Arbeit wurden zwei Klassifikatoren genutzt, zum einen Scikit-Learn (Gauß Naive Bayes), zum anderen die Libsvm Bibliothek, die mit einer Support Vector Machine arbeitet. Bei den Experimenten sind Klassifikationsfehler aufgetreten, die sich durch auffällige Ergebniswerte wie 800% Wahrscheinlichkeit bemerkbar gemacht haben. Das lag an teilweise sehr geringen Konfidenzwerten, die vom System beschnitten wurden und somit mathematische Formeln manipuliert haben. Generell war anfangs auch die Performance der Datenverarbeitung ein Problem. Für das Training wurden je 50% auf den beiden Korpora entnommen, der Rest wird für Validation verwendet. Die Ergebnisse stellen sich wie folgt dar: während Naive Bayes deutlich schneller arbeitet, liefert Libsvm präzisere Ergebnisse. Libsvm kommt auf 99% Accuracy, während Naive Bayes nur knapp über 50% liegt. Für Libsvm gab es zusätzlich die Unterscheidung zwischen Multi-Class und One-Class Varianten. Bei der One-Class Variante wird binär klassifiziert (0 oder 1). Obwohl Multi-Class noch bessere Ergebnisse liefert wäre die One-Class Variante vollkommen ausreichend für diese Arbeit gewesen. Weitere Metriken neben der Accuracy, die auf die Ergebnisse angwandt wurden sind Precision, Recall und der F1 Score. Im Großen und Ganzen lässt sich sagen, dass eine automatische Nachkorrektur der Ergebnisse sich als sehr sinnvoll erweist und sehr gute Ergebnisse liefert.

"Using morphologically rich POS-tagging to learn morphological generation"

Anastasia Kryvosheya, Betreuer: Dr. Alexander Fraser

Sprachen, die reich an morphologischer Flektierung sind, stellen eine besondere Schwierigkeit für computerlinguistische Bereiche wie Statistical Machine Translation dar. Um dieses Problem anzugehen, versucht Anastasia in ihrer Arbeit eine Art Wörterbuch zur Generierung solcher Flektierung zu erstellen, um solche Systeme zu unterstützen. Dabei wurden die Sprachen Russisch und Polnisch aufgrund ihrer Morphologie ausgewählt. Bei der maschinellen Übersetzung wird zunächst das Lemma von der Ausgangssprache in die Zielsprache übersetzt und dann wird mithilfe von morphologischen Eigenschaften-Tags die entsprechende Form generiert. Zwei Arten von Tools werden auf annotierten Korpora angewandt, um einen finalen, größeren Korpus zu gewinnen. Dabei ist das eine ein Lemmatizer, der Lemmata extrahiert, und das andere ein morphologischer Tagger, der die morphologische Information eines Wortes erfasst. Aus dem neu entstandenen, größeren Korpus wird ein Wörterbuch mit Frequenzen generiert, um zwischen den einzelnen Formen zu disambiguieren. Für die Generierung wird jeweils die häufigste Form verwendet. Anschließend sind noch einzelne Regeln im Einsatz, um das Ergebnis entsprechend zu korrigieren, wenn keine POS tags für ein Lemma vorhanden sind. Die getaggten Korpora, die verwendet wurden, sind zum einen der Russian National Corpus (1291k Tokens) und der

Polish National Corpus (126k Tokens). Zu den ungetaggten Korpora, die Teil dieser Arbeit waren, zählen der Yandex English-Russian Parallel Corpus (23271k Tokens) und der Europarl Parallel Corpus (7087k Tokens) für Polnisch. Dr. Schulz wirft ein, dass man Handarbeit in der Computerlinguistik keinesfalls unterschätzen sollte und Korpora nur einen kleinen Teil eines Wörterbuchs abbilden und somit keine ausführlichen und zufriedenenstellenden Wörterbücher generiert werden können. Der oben genannte Lemmatizer ist ein am CIS entwickeltes Tool names Lemming, der POS Tagger wird durch das Tool MarMOT repräsentiert. Diese beiden arbeiten an den ungetaggten Korpora, um diese mit Tags zu versehen. Dazu werden 80% der Daten aus den getaggten Korpora in das Training der Tools gespeist. Weitere 10% werden dem Developmentset zugeteilt und werden benötigt, um die Regeln für Fälle ohne POS Tag zu schreiben. Die restlichen 10% stellen das Testset dar. Anhand der Performance auf diesem Testset wird nun die Accuracy des Systems festgestellt. Die Ergebnisse für die polnische Sprache heben sich deutlich von den Ergebnissen für das Russische ab. Die polnische Accuracy erreicht 0.78% ohne den Einsatz von Regeln und 0.89% mit Regeln. Es zeichnet sich also eine Verbesserung von 10% ab, wenn man Regeln nutzt. Für das Russische hingegen fallen die Ergebnisse schlechter aus: 0.49% ohne Regeln und 0.53% mit Regeln, mit einer Verbesserung von nur 4%. Den Grund, den Anastasia für diese Diskrepanz zwischen den Sprachen nennt, ist dass der russische Korpus teilweise sehr fehlerhaft scheint. Die häufigste Flektionsform, die von den Tools ausgegeben werden soll, ist oft nicht korrekt. Das deutet darauf hin, dass die Tools Lemming und MarMOT Probleme bei der Verarbeitung russischer Korpora hatten. Anastasia schätzt die polnischen Ergebnisse als akzeptabel ein, wohingegen die russischen Ergebnisse zu schlecht seien. Daraus schließt sie, dass Preprocessing ein wichtiger Bestandteil von Natural Language Processing ist und nicht unterschätzt werden darf.