Katja Bertholdt
CIS, LMU München
20.06.17

Protokoll zur Sitzung vom 19.06.17 – Computerlinguistik Kolloquium

1. Vortrag: Ines Röhrer, „Musik und Ludwig Wittgenstein: Semantische Suche in seinem Nachlass“

Den ersten Vortrag des Tages hält Ines Röhrer. Ihrer Bachelorarbeit befasst sich mit dem Thema „Musik und Ludwig Wittgenstein: Semantische Suche in seinem Nachlass“ und wird von Dr. Hadersbeck betreut. Der Hauptfokus ihrer Arbeit liegt auf dem WITTFind-Tool und auf dem Nachlass Ludwig Wittgensteins. WITTFind ist eine vom CIS eigens für den Nachlass Wittgensteins konzipierte Suchmaschine mit zwei Suchoptionen: Einer regelbasierten Suche und einer semantischen Suche. Die semantische Suche enthält beispielsweise eine Kategorie „Farben“, dessen Implementierung Röhrer in ihrer Arbeit als Vorbild diente. Die Ergebnisse der semantische Suche werden in einer sog. Wordcloud dargestellt. Dabei handelt es sich um eine Darstellung, die besonders häufig vorkommende Wörter in einer größeren Schriftgröße darstellt als seltener auftretende Wörter.
Röhrers Ziel war es, eine Implementierung zu finden, die in das WITTFind Tool eingebunden werden kann und in der Applikation ein Modul für musikalische Begriffe zu ergänzen, wie es bereits für die Kategorie „Farben“ existiert. Des weiteren war die Untersuchung über Ontologien für Musikbegriffe eine Zielsetzung ihrer Arbeit (Inwiefern kann man diese Musikbegriffe als Ontologie modellieren? Welche Relationen existieren zwischen den Ausdrücken? Welche vorhandenen Tools kann man nutzen?).
Aus dem Nachlass Wittgensteins ist bisher nur ein kleiner Teil frei zugänglich, auf welchen sich WITTFind bisher beschränkt. Da aber viele musikalische Begriffe in einem deutlich größeren, noch nicht veröffentlichten Teil zu finden sind, bezieht Röhrer auch diesen in ihrer Arbeit mit ein.
Die Motivation für ihre Arbeit liegt darin, dass das Leben Wittgensteins stark von Musik geprägt war und deshalb ein Interesse daran besteht, wie er diese in seinem Nachlass verarbeitet und erwähnt.
Ein wichtiger Teil der Umsetzung bestand in der Erweiterung des Webfrontends. Dazu hat Röhrer die entsprechenden Dateien für die Ergänzung ihres Moduls erweitert.
Als Basis ihrer Arbeit diente eine Hausarbeit eines Musikwissenschaftlers der LMU. Anhand dieser Hausarbeit hat Röhrer sechs Kategorien für ihre semantische Suche festgelegt: „Komponisten“, „Gattungen“, „Instrumente“, „Intervalle“, „Bezug zu Komposition“ und „Sonstige Begriffe“.
Um eine Darstellung durch Wordclouds ermöglichen zu können, müssen Frequenzberechnungen erfolgen. Dazu hat Röhrer ein Dictionary erstellt, mit den musikalischen Wörtern als key und der dazugehörigen Herkunftsdatei und der Frequenz als value. Da für eine korrekte Frequenz die Vollformenextraktion wichtig ist, hat Röhrer zunächst eine von ihr händisch erstellte Endungsliste verwendet. Später hat sie ein Vollformenlexikon für diesen Zweck benutzt. Da im Lexikon jedoch einige flektierte Vollformen nicht ihrem Lemma zugewiesen sind, wurden mit der Endungsliste bessere Ergebnisse erzielt.
Im Laufe Ihrer Arbeit ist die Betrachtung des Kontexts der einzelnen Wörter stärker in den Vordergrund gerückt. Deshalb hat sich Röhrer mit der Kontextextraktion der Wörter beschäftigt. Sie hat zwei unterschiedliche Kontextvarianten betrachtet: zum einen mit Einbezug der Stoppwörter, zum anderen ohne Stoppwörter. Für die Extraktion hat sie zwei verschiedene Methoden angewandt: einen Ringbuffer und Listenoperationen. Während beim Ringbuffer für jede Kontextvariante ein eigener Ringbuffer verwendet werden muss, können bei der Listenoperation beide Varianten gleichzeitig berücksichtigt werden, weshalb Röhrer diese Variante bevorzugt. Dabei werden die Teile („Bemerkungen“) des Nachlasses als Liste eingelesen und durch eine Abfrage die Relevanz

eines Wortes geprüft und dessen Kontext mit bzw. ohne Berücksichtigung der Stoppwörter der Liste entnommen.
Durch Verschiebung des Fokus auf die Kontexte, blieb Röhrer weniger Zeit für die Auseinandersetzung mit Ontologien. Röhrer hatte zum Ziel, die Relationen zwischen musikalischen Begriffen auf der Website zur Verfügung zu stellen. Diese Aufgabe stellte sich als eine komplexes Problem heraus. Jedoch ist Röhrer auf eine Musikontologie („The Music Ontology“) gestoßen, mit der ihr eine beispielhafte Modellierung einiger Komponisten gelang. Auf eine weitere Ausarbeitung musste sie aufgrund zeitlicher Probleme verzichten.
Abschließend fasst Röhrer zusammen, dass ihre Arbeit erfolgreich war, auch wenn sie ihre Ergebnisse und Vorgehensweisen noch für verbesserungswürdig hält. Außerdem betont sie die große Interdisziplinarität ihrer Arbeit, die große Teile der Philosophie und Musik beinhaltet.

2. Vortrag: Michael Strohmeyer, „Machine-Learning basierte automatische OCR-Korrektur“

Den zweiten Vortrag hält Michael Strohmeyer. Er stellt das Thema seiner Bachelorarbeit vor, in der er sich mit Machine-Learning basierter automatischer OCR-Korrektur beschäftigt hat. Dabei wurde er von Dr. Schulz betreut. Die Motivation der Arbeit besteht darin, dass sich Schriftbilder und Schreibweisen der Wörter mit der Zeit verändern. Bei der Digitalisierung durch OCR-Systeme (optical character recognition) werden deshalb manche Wörter nicht zuverlässig erkannt. Wenn sich das OCR-System bei einem Wort in der Erkennung nicht sicher ist, liefert es Korrekturvorschläge für dieses Wort. Als Beispiel nennt Strohmeyer die Wörter „Hund“ und „Hand“. Das Ziel Strohmeyers Arbeit bestand darin, eine Software zur automatischen Nachkorrektur der eingelesenen OCR- Dokumente zu erstellen. Dazu sollte ein Machine Learning System trainiert werden.
Strohmeyer hat in seiner Arbeit zwei Grund-Truth Dokumente verwendet. Zum einen „Paradiesgärtlein“ und zum anderen „Curiöser Botanicus“. Beide stammen aus dem RIDGES Korpus, der am CIS der LMU in Kooperation mit der Humboldt Universität Berlin erstellt wurde und 33 Kräuterkundetexte aus der Zeit zwischen 1484 und 1914 beinhaltet.
Strohmeyer hat zunächst die ihm zu Verfügung stehenden Dokumente eingelesen. Danach hat er die Feature-Werte importiert und angewendet, die ihm vom Profiler geliefert wurden. Ein Profiler liefert gewichtete Interpretationen zu allen Wörtern, die in der OCR-Ausgabe stehen. Diese Interpretationen beziehen sich einerseits auf Erkennungsfehler und sagen aus, wie das Wort im ursprünglichen Text heißen müsste. Andererseits wird die Schreibweise berücksichtigt und es werden Aussagen darüber getroffen, wie das „moderne“ Wort inzwischen geschrieben wird (zB. „Thon“ und „Thon“). An dieser Stelle zeigt Strohmeyer den Aufbau der Profiler-Ausgabe anhand eines Beispiels. Daraus hat er einige Werte als „Basis-Features“ extrahiert: den Konfidenzwert (Prozentzahl, die angibt wie sicher sich der Profiler bei seiner Interpretation ist), die Levenshtein-Distanz (wie groß ist die Abweichung von OCR-Ausgabe und Korrekturvorschlag?) und die Häufigkeiten. Danach hat er zusätzliche Features ergänzt: die Längendifferenz zwischen OCR-Ausgabe und Korrekturvorschlag und den nächsthöheren Konfidenzwert der Korrekturvorschläge.
Als Machine Learning Klassifikatoren hat Strohmeyer den Naive Bayes Klassifikator der Scikit-learn Bibliothek und Libsvm verwendet. Der Naive Bayes Klassifikator trifft Aussagen mithilfe von Wahrscheinlichkeitsrechnungen, während Libsvm eine Support Vector Machine zur Klassifizierung von Daten verwendet.
In seiner Arbeit ist Strohmeyer zunächst auf Performance Probleme in der Datenverarbeitung gestoßen. Diese ließen sich durch eine Änderung der internen Datenstruktur beheben. Außerdem wurden in der Ausgabe des Profilers die unterste Zeile nicht vollständig angezeigt, sodass die daraus als Features entnommenen Werte zu einer Fehlklassifikation geführt haben. Deshalb hat Strohmeyer die benötigten Werte aus der Kommentarzeile des Profilers ausgelesen, die die identischen Werte beinhaltet und vollständig lesbar war.
Mittels Kreuzevaluierung wurden in Strohmeyers Arbeit sehr gute Ergebnisse erzielt. Dazu wurden 50% des einen Korpus und 50% des anderen Korpus verwendet, getestet und evaluiert. Beim

Vergleich der beiden verwendeten Klassifikatoren hat Strohmeyer festgestellt, dass Naive Bayes zwar deutlich schneller war, Libsvm jedoch bessere Ergebnisse geliefert hat. Er hat außerdem drei verschiedene Kombinationen der Features getestet und verglichen: im ersten Fall wurden nur die Basisfeatures verwendet, im zweiten die Basisfeatures und der nächsthöhere Konfidenzwert und im letzten Fall die Basis-Features und alle ergänzten Features. Dabei hat der dritte Fall am besten abgeschnitten.
Strohmeyer hält die automatische Nachkorrektur von OCR Dokumenten für sinnvoll, da es gute Ergebnisse liefert. Als Ausblick für mögliche Verbesserungen nennt er eine Kombinierung der beiden Klassifikatoren oder die Ergänzung weiterer Features.

3. Vortrag: Anastasiya Kryvosheya, „Using morphologically-rich POS tagging to learn morphological generation“

Den letzten Vortrag hält Anastasiy Kryvosheya. Ihre Bachelorarbeit behandelt das Thema „Using morphologically-rich POS tagging to learn morphological generation“ und sie wird von Dr. Fraser betreut. Die Motivation ihrer Arbeit besteht darin, dass es Sprachen gibt die eine reiche Morphologie aufweisen und deshalb eine Herausforderung für viele Bereiche der Computerlinguistik darstellen. Für diese Sprachen ist ein komplexes Flektierungssystem charakteristisch. In ihrer Arbeit hat Kryvosheya dafür das Polnische und die russische Sprache als Repräsentanten dieser Sprachen ausgewählt. Im Polnischen gibt es sieben und im Russischen sechs Kasus. Im Gegensatz zur deutschen oder englischen Sprache wird in diesen Sprachen mittels der Flektierung grammatische Informationen des Wortes vermittelt (im Deutschen ist dies lediglich beim Genitiv-s und in der Pluralbildung zu finden).
In den letzten Jahren hat die SMT (statistical machine translation) immer mehr an Bedeutung gewonnen. Diese Übersetzung besteht aus zwei Schritten. Im ersten Schritt wird Wort in ein Lemma übersetzt und seine morphologischen Features herausgefunden. Im zweiten Schritt wird dann aus dem Lemma die Form entsprechend der gespeicherten Features generiert. Für diesen zweiten Schritt wird morphological generation benötigt und somit handelt es sich dabei um ein Subtask von SMT.
Das Ziel ihrer Arbeit war es, mithilfe eines getaggten Korpus ein Generierungssystem aufzubauen, das für jedes Wort und seine morphologischen Eigenschaften eine Form generieren kann.
Zunächst hat Kryvosheya mithilfe eines Lemmatizers und eines morphologischen Taggers ein Modell auf einem annotierten Korpus trainiert. Durch dieses wurde dann unannotierten Korpora getaggt, um insgesamt einen größeren annotierten Korpus zu erhalten. Aus diesem getaggten Korpus wurde ein Wörterbuch mit POStag, morphologischen Eigenschaften und Häufigkeit erstellt. Um zwischen den Formen zu disambiguieren, wurde die am häufigsten vorkommende Form als die richtige deklariert, da im getaggten Korpus beispielsweise häufig eine morphologische Eigenschaft zwei unterschiedliche Formen hatte. Für Wörter, deren POStag und morphologische Eigenschaften nicht im Wörterbuch gefunden wurden, hat Kryvosheya versucht, grammatische Regeln für die Generation anzuwenden.
Kryvosheya wollte in ihrer Arbeit prüfen, ob es möglich ist, nur anhand eines getaggten Korpus den Schritt der morpholgical generation in SMT durchzuführen, da viele andere Tools eine aufwändigere Vorgehensweise verwenden.
Als getaggte Korpora hat Kryvosheya für das Russische den Russian National Corpus und für das Polnische den Polish National Corpus verwendet. Als ungetaggte Korpora hat sie für das Russische den Yandex English-Russian Parallel Corpus und für das Polnische den Europarl Parallel Corpus (Poish) eingesetzt. Als Lemmatizer wurde „Lemming“ verwendet, ein Lemmatizerprogramm, das am CIS entwickelt wurde. Außerdem wurde das Programm „MarMOT“ verwendet, um die Wörter mit dem entsprechenden POStag und ihren morphologischen Eigenschaften zu taggen.
Es wurden 80% der getaggten Korpora Lemming und MarMOT übergeben, um ein Modell zu erstellen, mit dem dann die ungetaggten Korpora getaggt wurden. Die Accuracy für das Polnische betrug bei der Generierung ohne Regeln 78% und mit Regeln 89%. Im Russischen wurde ohne

Regeln eine Accuracy von 49% und mit Regeln 53% erzielt. Die schlechteren Ergebnisse des Russischen lassen sich laut Kryvosheya damit erklären, dass „Lemming“ und „MarMOT“ auf dem ungetaggten russischen Korpus falsche Ergebnisse geliefert haben.
Für die häufigsten Fälle, für die kein Lemma generiert werden konnte, hat Kryvosheya versucht, Regeln zu entwickeln. Obwohl sich diese Idee als schwierig herausgestellt hat, wurden 797 Formen dadurch richtig generiert.
Abschließend betont Kryvosheya erneut die Schwierigkeiten beim taggen des ungetaggten russischen Korpus und erwähnt, dass die festgelegten Regeln zu einer Verbesserung der Accuracy bei der Generierung in der polnischen Sprache um 10% beigetragen haben und in der russischen Sprache um 3%.