CIS, LMU München
22.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 22.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Faridis Alberteris, Julia Khobotova und Alexander Vordermaier jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Optimierung der linguistischen Suche beim XML-annotierten Nachlass von Ludwig Wittgenstein

Das erste Referat wird gehalten von Faridis Alberteris Azar, betreut von Herrn Hadersbeck. Ihr Thema ist die Optimierung der linguistischen Suche beim XML-annotierten Nachlass von L. Wittgenstein.
Im Rahmen von Digital Humanities Projekt Wittgenstein in Co-Text wurde in Zusammenarbeit mit der Universität Bergen in Norwegen eine Finder App „WiTTFind“ entwickelt. Die Universität Bergen hält die Rechte von 20000 Manuskripten aus Wittgensteins Nachlass. 5000 annotierte Seiten daraus wurden dem CIS zu EDU Zwecken überlassen.
Das Ziel der Bachelorarbeit soll es sein, durch die bestmögliche Ausnutzung der XML Tags eine Verbesserung der XML Annotation zu erreichen. Schwerpunkt wird hier auf die Personensuche gelegt.
Nachdem die Referentin ihr Thema erklärt hat, zeigt sie den grundsätzlichen Aufbau der von ihr bearbeiteten Dateien. Diese gliedern sich in ORG (Originaldatei), NORM (normalisierte Datei) und DIPLO (diplomatische Datei = mögliche Änderung nicht final). Die ORG Dateien haben hierbei keine XML Struktur. Durch taggen wird eine NORM.xml generiert. Dafür wird ein probabilistischer, auf Markowmodellen basierender Part-Of-Speech Taggger von Helmut Schmid verwendet.
Auf dieser Basis werden Eigennamen lokalisiert (Schritt 1). Hierbei wird außerdem überprüft, welche Fehler beim Tagging gemacht werden. Im 2. Schritt soll dann die Suche in WittFind verbessert werden, indem eine neue Kategorie in WittFind und im CIS Lexikon eingetragen wird.
Abschließend kommt die Referentin zu ihrer Evaluierung. Im ursprünglichen System wurden in 13 zur Verfügung stehenden Dateien insgesamt 168 Personen gefunden. Dem gegenüber stehen 833 gefundene Personen im neuen, empfohlenen System. Diese wurden händisch auf Richtigkeit überprüft und weisen eine hohe Precision-Rate auf.

CIS, LMU München
22.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 22.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Faridis Alberteris, Julia Khobotova und Alexander Vordermaier jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Comparison of Transfer Methods for low Ressource Morphology

Das letzte Referat der Sitzung hält Alexander Vordermaier. Er hat das Thema Compaarison of Transfer Methods for low Ressource Morphology bei Katharina Kann gewählt.
Der Kern seiner Arbeit ist die Paradigmen Komplettierung von Sprachen, also die Zuordnung eines Lemmas zu seinen flektierten Formen. Hierbei tritt das Problem auf, dass bei Low Ressource Sprachen nur wenig Quellen vorhanden sind. Deshalb wird eine ähnliche High Ressource Sprache zum Umgehen des Problems hinzugezogen. In seinem Fall war die LR Sprache Mazedonisch und die HR Sprache Bulgarisch. Für die Paradigmen Komplettierung werden drei Methoden verwendet: Sprachübergreifende Paradigmen Kompl. , Auto Encoding und eine Kombination aus den beiden Modellen.
Seine Herangehensweise war die Vermischung der annotierten Daten der LR mit den annotierten Daten der HR Sprache. Beim Autoencoding wurde ein sehr simples Verfahren verwendet, bei dem die Eingabe gleich der Ausgabe ist. Hierbei hofft man, dass die Flektionen der Wörter der beiden Sprachen in etwa gleich ist, geht jedoch ein hohes Risiko ein, dass dies nicht der Fall ist. Diese Daten wurden dann vom Modell trainiert., indem sie in unterschiedliche Paketgrößen aufgeteilt und dann immer paarweise vermischt wurden.
Anschließend präsentierte Alexander noch die Form der Ein- und Ausgabe Dateien und in welchem Format seine Daten vorlagen anhand von Beispielen.

Als letztes gab der Referent noch einen Ausblick auf seine Ergebnisse, die Fehleranalyse und die weiteren Aufgaben.
Er kam zu dem Schluss, dass leider sehr oft die falsche Endung verwendet wird und beim Autoencoding generell sehr viele Fehler aufgetreten sind. Dies hatte er sich schon gedacht, da es auf Grund der Vorgehensweise zu erwarten war. Außerdem war die Evaluation für ihn nicht einfach, da er beide verwendeten Sprachen nicht beherrscht und somit nicht immer genau wusste, wo die Fehler gemacht wurden. Dies sieht er auch als eine seiner weiteren Aufgaben; die Identifizierung weiterer Fehlerquellen. Außerdem will er noch bereits gängige Verfahren beleuchten.