CIS, LMU München

Studentin: Elena Atanasova

22.05.2017

Protokoll zur Sitzung vom 22.05.2017 – Computerlinguistisches Arbeiten

1. Vortrag: Faridis Alberteris Azar, "Optimierung der linguistischen Suche beim XML-annotierten Nachlass von Ludwig Wittgenstein"

BA-Betreuer: Dr. Maximilian Hadersbeck

Den ersten Vortrag des heutigen Tages hält Alberteris Azar, die von Dr. Hadersbeck betreut wird. Sie schreibt ihre Bachelorarbeit im Rahmen des Digital-HumanitiesProjekts „Wittgenstein in Co-Text", im Zusammenarbeit mit dem Wittgenstein Archivs der Universität Bergen (WAB), in Norwegen (Typescript und Manuscript ). Seit 2010 kooperieren WAB und das Centrum für Informations- und Sprachverarbeitung (CIS) in der Forschungsgruppe: "Wittgenstein in Co-Text". Die Gruppe entwickelte die web-basierte FinderApp WiTTFind. Die Aufgabe von Frau Alberteris Azar ist die Linguistische Suche in WiTTFind zu verbessern, Transkriptionsfehler und Editionsprobleme erkennen und so viel wie möglich lösen und Erweiterung des Lexikons.

Das Ziel ihre Arbeit ist die Optimierung der linguistischen Suche beim XML-annotierten Nachlass von Ludwig Wittgenstein durch die optimale Ausnutzung der XMLAnnotation und die Verbesserung der XML-Edition mit Schwerpunkt auf Personennamen und Erweiterung des Lexikons.

Die XML-annotierten Editionstexten stammen aus den XSLT-Dateien aus Bergen, die für die Konvertierung der originalen XML-Editionen dienen: ORG stehht für alle Optionen (sie hat uns das Beispiel [Eine herrenlose Wohnung] gezeigt), NORM ist für was es sein sollte und DIPLO wie Ludwig Wittgenstein es geschrieben hat.

Sie verwendet ein TreeTagger, der von dem Herr Dr. Helmut Schmid (CIS Professor) entwickelt wurde, basiert sich auf dem Marko Model und unterscheidet sich durch die Entscheidungsbäume für das Messen für Übergangswahrscheinlichkeiten.

Für die Eigennamen- Erkennung benutzte Frau Alberteris Azar die alte NORM- Dateien. Der Tagger taggt die NORM- Dateien und generiert die NORM-tagged.xml Dateien. Im Februar hat (CIS) aktualisierte Dateien aus Bergen bekommen mit einem neuen XML-Element und zwar persNamen. Damit wird die Arbeit viel erleichtert.

Als nächstes hat Frau Alberteris Azar die Resultate ihrer Arbeit aus 20 Dateien presentiert. WittFind findet jetzt in 13 Dateien insgesamt 168 Treffer. Das empfohlene System findet in alle Dateien insgesamt 833 Treffer.

Sie hat schon Ideen wie bessere Ergebnisse zu bekommen. Die sind: neue syntaktische Kategorie „persName" zu erzeugen. Als Mittel etree zu benutzen und in getaggte Dateien bei jeden gesammelten Beispielen „persName" hinzufügen. Neue Kategorie in WittFind erzeugen und Neue Kategorie in CIS-Lexikon bei EN eintragen.

z.B. Russellschen,Russell.EN+persName

Russellsche,Russell.EN+persName

Russells,Russell.EN+persName:geM

Frau Alberteris Azar wird auf Verbesserung der Tokenisierung, Verbesserung des Tagging, Verbesserung der Personen Namenerkennung, Erweiterung des Lexikons, wenn man die Wortliste bzw. die Frequenzliste mit Hilfe von etree anstatt Regex erzeugt, arbeiten.