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

Faridis hat den ersten Vortrag des Tages gehalten. Ihr Betreuer ist Dr. Maximilian Hadersbeck. Sie schreibt ihre Arbeit im Rahmen des Digital Humanities Projekt und in Zusammenarbeit mit dem Wittgenstein Archiv der Uni Bergen in Norwegen. In dieser Zusammenarbeit wurde bereits am CIS die Suchmaschine WITTFIND entwickelt. Damit kann man zum Beispiel nach Wörtern oder Begriffen im Nachlass von Ludwig Wittgenstein suchen. Die Manuskripte aus dem Nachlass wurden von der Uni Bergen in XML Dateien transkribiert. Ziel der Arbeit ist die Optimierung von WITTFIND, mit dem Schwerpunkt auf Personennamen der Erweiterung des Lexikons.

Dabei wurden die Originaldateien speziell für das CIS angepasst. Zu jeder Seite gibt es drei verschiedene XML-Dateien: das Original, die Normalisierung und die Diplomatische(?) Version.

In den originalen Manuskripten hat Wittgenstein oft Verbesserungen eingefügt. Die Originaldateien enthalten also alle Möglichkeiten die sich mit diesen Verbesserungen ergeben. Dagegen wurden in den Diplomatischen Dateien keine großen Änderungen vorgenommen. In der Arbeit konzentriert man sich aber nur auf die normalisierten Dateien, dort steht alles so darin, wie es eigentlich seien sollte.

In der Suchmaschine wird der Probabalistischer POS Tagger „TreeTagger“ verwendet. Wie der Name sagt, verwendet der Tagger Entscheidungsbäume zum Messen von Übergangswahrscheinlichkeiten. Dieser wird dann bei den normalisierten Dateien angewendet.

Der Tagger hatte aber noch einige Probleme mit der Erkennung von Eigennamen und somit will man in der Bachelorarbeit Vorschläge machen, wie man die XML Informationen besser verwenden kann. Dies passiert in zwei Schritten.

Schritt 1: Man lokalisiert alle Beispiele, in denen der Tagger sich falsch verhält. Zusätzlich hat man mit dem Python Parser „Etree“ pro Dokument eine Liste mit allen Personennamen erstellt.

Schritt 2: Man will die semantische Suche in WITTFIND verbessern

Bei den Resultaten nannte sie eine Trefferquote von 168 Treffern bei der Namensbildung. Wenn die Vorschläge bedacht werden würden, würde man auf eine Trefferquote von 833 kommen.

Zum Schluss hat sie weiteren Aufgaben genannt. Darunter fallen bessere Tokenesierung und Tagging und die Erweiterung des Lexikons