Klyueva Olga
CIS, LMU München
22.05.1917

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

Die Bachelorarbeit wird von Herrn Dr. Maximilan Hadersbeck betreut.
Das Ziel der Bachelorarbeit ist eine Optimierung der Linguistischen Suche bei dem 5000 Seiten umfassenden XML-annotierten Nachlass von Ludwig Wittgenstein durch die optimale Ausnutzung der XML-Annotationen und Verbesserung von XML-Additionen. Der Schwerpunkt wurde auf Personennamen gesetzt und dadurch wurde auch das Lexikon erweitert. Es gab Probleme in der Personennamenerkennung: nicht alle Personennamen wurden identifiziert, einige wurden falsch erkannt.

Als erstes wurde über eine Suchmaschine Wittfind erzählt, die im CIS entwickelt wurde. Wittfind sucht nach Wörtern und Begriffen in den Typescripten und Manuskripten von Ludwig Wittgenstein. Typescripten- und Manuskripten-Sammlung heißen Nachlass und wurden von Ludwig Wittgenstein maschinell oder mit Hand geschrieben. Der Nachlass wurde in XML-Dateien transkribiert und 5000 von 20 000 XML Seiten nach CIS geschickt. Wittfind arbeitet mit Dateien mit 5000 XML Seiten.
Es wurden drei Typen XML-Dateien für CIS erzeugt: ORG (Originaldateien), Norm (Normalisierte Dateien), Diplo (Diplomatische Dateien).
ORG XML Dateien sind digitalisierte Dateien mit allem, was Wittgenstein in seine Texte eingefügt, durchgestrichen oder gelassen hat. Diplomatische Version ist eine Version, bei der er die Wörter in Manuskripten oder in Typeskripten gelassen hat. Normdatei ist eine normalisierte Version. Im Rahmen der Bachelorarbeit wurde nur mit normalisierten XML-Dateien gearbeitet.

In der Suchmaschine Wittfind wurde ein probalistischer POS-Tagger (ein Treetagger), entwickelt von Herrn Dr. Helmut Schmid, verwendet. Der POS-Tagger basiert auf Markov Modelle. Treetagger läuft durch Norm XML Dateien, tokenisiert und taggt alle Wörter.

Als nächster Punkt der Präsentation wurde Eigennamenerkennung genannt. Es wurde ein Stück von Norm-XML-Dateien präsentiert und erklärt, wie der Tagger sich verhält. Der Tagger läuft durch eine Norm-Datei und erzeugt ein neues Dokument mit dem neuen XML-Element, das wt (wort) heißt. Außerdem es gibt zwei Attribute: Attribut Tag, Attribut Lemma. Die Personennamen wurden an einigen Stellen falsch getaggt, beispielweise Adjektive als Personennamen getaggt.
In der Bachelorarbeit versucht man Vorschläge zu geben, wie man Information von XML Dateien besser verwenden kann, damit Wittfind zu besseren Resultaten in Personennamenerkennung kommt.

Es wurden einige Schritte erwähnt.

1. Lokalisierung der Fehler. Außer POS-tagger wurde etree (Python Parser) verwendet, der Personennamen in XML Dateien erkennt.
2. Die Verbesserung der semantischen Suche in Wittfind. In Wittfind werden Eigennamen nach einem Muster gesucht. Es sind Fehler aufgetreten. Wittfind bezeichnet das Wort “Venus“ oder „hellerische“ als Personennamen, obwohl es nicht so ist. Zur Verbesserung der Suche wurde ein syntaktisches Attribut zu Suchmuster <PersNamen> hinzugefügt. Dadurch werden aus 20 Dateien 833 Personennamen erkannt. Vorher wurden in 13 aus 20 Dateien 168 Treffer und eigene Personennamen falsch erkannt.

Zum Schluss wurden weitere Kapitel der Bachelorarbeit erwähnt: Transkriptionsfehler vs. Editionsprobleme bei XML-Dateien in Wittgensteins Nachlass. Es wurde die Verbesserungsschritte genannt: Verbesserung der Tokenisierung, Verbesserung des Tagging und Verbesserung der Personennamenerkennung.

Weiterer Schritt ist eine Erweiterung des Lexikons mit Personennamen, wenn man die Wortliste beispielsweise die Frequenzliste mit Hilfe von Etree anstatt Regex erzeugt.