Studienprojekt von Simone Eberhard und Katja Niemann
im Rahmen des Projekts „Semantisches Information Retrieval“ (SIR)
Betreuerin: Dr. Iryna Gurevych

Projektbeschreibung:

Das Projekt SIR richtet sich auf die Entwicklung von Methoden des Information Retrieval (IR), die über den gegenwärtigen Stand der Technik hinausgehen. Im Gegensatz zu den momentan eingesetzten Ansätzen, wird die Relevanz eines Dokuments, in Bezug auf eine Anfrage, mit Hilfe von Maßen der lexikalisch-semantischen Ähnlichkeit bestimmt.

Aufgabenbeschreibung:

Unser Studienprojekt hat eine Verbesserung und Erweiterung des SIR-Projektes zum Ziel. Wir implementieren ein neues Maß semantischer Ähnlichkeit in einer abgewandelten Form nach Leacock und Chodorow, um die semantische Verwandtschaft von Wortpaaren, die aus einem Text extrahiert wurden, zu berechnen.

Diese Wortpaare müssen vorverarbeitet werden, dazu wird der Smes-Stemmer sowie der TreeTagger in SIR integriert. Zudem müssen Komposita vor der Berechnung getrennt und die einzelnen Verwandtschaften in geeigneter Weise addiert werden.

Für das Maß von Leacock und Chodorow muß der kürzeste Pfad zwischen zwei gegebenen Wörten, sowie der längste Pfad in der Taxonomie berechnet werden.
Da das Programm später auf einem sehr großen Korpus laufen soll, ist das Laufzeitverhalten von großer Bedeutung.

Die Implementierung erfolgt in Java-Klassen und Paketen, sowie shell-scripts zum ausführen auf der LinuxPlattform.

Wenn unser Maß in das SIR-Projekt integriert ist, werden wir noch eine graphische Benutzeroberfläche erstellen.