Satzerkennung

Betreuer:

Peter Siniakov: siniakov@inf.fu-berlin.de

Zielgruppe:

Das Projekt richtet sich an Studenten im Hauptstudium, die Interesse an der Verarbeitung natürlicher Sprache haben. Linguistische Vorkenntnisse sind vom Vorteil, werden jedoch nicht zwingend vorausgesetzt. Kenntnisse der Grundtechniken der KI sind sehr wünschenswert.

Motivation:

Die Erkennung der Grenzen eines Satzes in einem Text wäre eine triviale Aufgabe, wenn die Satzzeichen eine eindeutige Semantik hätten. Da sie jedoch oft überladen sind, wird die Erkennung erheblich erschwert. Bis heute gibt es noch kein linguistisches Werkzeug für die deutsche Sprache, das die Grenzen eines Satzes mit sehr hoher Zuverlässigkeit erkennen kann.
Die zuverlässige Satzgrenzenerkennung ist jedoch für viele Felder der Sprachverarbeitung eine wichtige Voraussetzung. Beim syntaktischen Parsing kann bei einer falschen Erkennung der Satzgrenzen die Zuordnung der syntaktischen Funktionen scheitern. Für die Gebiete der Informationsextraktion und Textverständnis trägt ein Satz als Textelement eine semantische Information. Ein Satz fasst gewöhnlich zusammenhängende Inhalte zusammen und drückt einen abgeschlossenen Gedanken aus. Oft ist er auch eine grundlegende Struktur für den Ausdruck eines Fakts und stellt somit ein semantisches Strukturierungselement dar.
Die Ergebnisse des Projekts werden in ein System zur Faktenextraktion aus natürlich-sprachlichen Texten einfließen, das im Rahmen des Forschungsprojekts FEx in der Arbeitsgruppe Datenbanken und Informationssysteme entwickelt wird.

Ziele des Projekts:

Die Satzgrenzen in einem natürlich-sprachlichen Text sollen gefunden und gekennzeichnet werden. Alle vorkommenden Nebensätze sollen erfasst und typisiert werden.

Durchführung:

Am Beginn des Projektes sollen die Teilnehmer sich in den aktuellen Forschungsstand vertiefen. Insbesondere sollen die Ansätze für die syntaktische Analyse kritisch betrachtet und hinsichtlich ihrer Tauglichkeit und Realisierbarkeit bewertet werden. Vorhandene linguistische Werkzeuge (morphologische Analyse, syntaktischer Parser) können zum Aufbau eines Trainingscorpus benutzt werden. Das entstehende Programm muss als eigenständiges Modul funktionieren, das einen Text mit semantischen Informationen über Satz- und Nebensatzgrenzen sowie den Typ der Nebensätze in Form von XML-Annotationen anreichert. Abhängig von der Qualität der Ergebnisse soll ein Algorithmus für die Korrektur der von anderen linguistischen Werkzeugen falsch erkannten Satzgrenzen entwickelt werden.

Das Projekt soll von einer kleinen Gruppe durchgeführt werden. Start und Ende des Projektes können mit dem Betreuer flexibel vereinbart werden, genauso wie die regelmäßige Projekttreffen. Das Projekt beginnt mit einer kleinen Vortragsreihe, in der die Projektteilnehmer kurze Vorträge über den aktuellen Forschungsstand halten. Danach erfolgt die Planung, Design und Implementierung des Tools.