Karlsruher Institut für Technologie

Fakultät für Informatik
Institut für Programmstrukturen und Datenorganisation
Prof. Walter F. Tichy
Internetadresse: http://www.ipd.uka.de/Tichy

BA

Optimieren von POS-Tagger-Ergebnissen

Ausgangssituation/Kontext

Im Forschungsbereich „Sprachverarbeitung in der Softwaretechnik“ kommen Werkzeuge von Computerlinguisten zum Einsatz. Ein grundlegendes Werkzeug ist ein Part-of-Speech-Tagger (POS-Tagger), welcher zu den Wörtern eines gegebenen Textes die Wortarten hinzufügt, z.B. **Der**[Artikel] **Ball**[Substantiv] **ist**[Verb] **rund**[Adjektiv]**.** Diese Analyse ist oft Grundlage weiterer linguistischer Analysen.
POS-Tagger werden meist auf Texten einer bestimmten Quelle trainiert und entsprechend getestet. Leider bereitet dieses Training die Tagger nicht auf den speziellen Einsatz für Texte aus der Softwaretechnik vor.

Ziel

Diese Arbeit soll bestehende POS-Tagger und deren unterschiedlichen Konfigurationen auf Ihre Einsetzbarkeit in unserem Kontext untersuchen. Hierzu soll ein automatisches Vergleichsverfahren basierend auf einem Benchmark implementiert werden.
Die Arbeit soll außerdem die Frage beantworten, ob ein Training der POS-Tagger auf unseren Texten eine Verbesserung bringen kann und/oder ob die Ergebnisse verschiedener Tagger (oder Konfigurationen) zu einem besseren Ergebnis führt.

Voraussetzungen

Für diese Arbeit bringen Sie Spaß am Umgang mit natürlicher Sprache mit; um für die Implementierung gerüstet zu sein, verfügen Sie über Programmiererfahrung (vorzugsweise in Java). Sie haben keine Angst davor, im Team zu Arbeiten und scheuen sich nicht, neue, Ihnen unbekannte Techniken einzusetzen. Außerdem zögern Sie nicht, eine E-Mail zur Vereinbarung eines ersten Gesprächs zu schreiben, in welchem wir Ihnen unverbindlich Details und einen persönlichen Eindruck unserer Arbeit geben werden.

Informatikerfreundliche Arbeitsumgebung

- Redundante Kaffeemaschinenanbindung
- Klimatisierter Poolraum
- Gut ausgebaute Süßigkeiteninfrastruktur

Betreuer

Mathias Landhäußer, Raum 343
Sprechzeiten nach Abstimmung, landhaeusser@kit.edu
Sebastian Weigelt, Raum 346
Sprechzeiten nach Abstimmung, sebastian.weigelt@partner.kit.edu