Dr.-Ing. Hui Shi
FB3 Mathematik und Informatik
Universität Bremen

Telefon (0421) 218 - 64260
Fax (0421) 218 – 9864260
shi@informatik.uni-bremen.de

Postfach 33 04 40 · 28334 Bremen Germany

Bremen, den 30.01.2015

Gutachten

über die Bachelorarbeit

Automatische Generierung einer Wissensbasis mithilfe mehrsprachiger Korpora

vorgelegt von Herrn Tim Nieradzik

Automatische Generierung von Wissensgrundlagen ist ein Thema mit großen Herausforderungen sowohl in Theorie und auch in Praxis in der Computerlinguistik. Herr Tim Nieradzik nimmt sich diese Herausforderung an und beschäftigt sich in seiner Bachelorarbeit mit einer Teilaufgabe des Themengebiets, nämlich der Klassifizierung der deutschen Satzbestandteile mithilfe einer anderen Sprache (hier Polnisch). Seine Untersuchungsergebnisse zeigen, dass durch Hinzunahme einer zweiten Sprache die Mehrdeutigkeit der Klassifizierung reduziert werden kann.

Die Arbeit teilt sich neben dem Einleitungskapitel und der abschließenden Diskussion in 3 Kapitel auf.

Das zweite Kapitel dient zur Erläuterung der linguistischen Begriffe (wie Morphem, Token und Kompositum) und Methoden (wie Tagging und morphologische Analyse), wobei der Unterschied von „Lexem“ und „Token“, die in der Arbeit häufig verwendet sind, unklar bleibt. Die bekannten Korpora und Wörterbücher für Deutsch und Polnisch werden in §2.2 und §2.3 präsentiert und das Korpus OpenSubtitels 2013 wird für die vorgelegte Bachelorarbeit ausgewählt. In §2.4 stellt der Autor zahlreiche existierende Tagger vor, anschließend wird die morphologische Analyse, als eine Erweiterung einer existierenden Analyse, präsentiert. Wie die Analyse für die weitere Arbeit benutzt wird, wird nicht beschrieben.

Im Kapitel 3 wird zuerst der Begriff „semantische Einheit“ definiert, um die hierarchische Annotation von Sätzen und semantischen Abbildungen zwischen Sätzen in verschiedenen Sprachen zu unterstützen. Die bewehrten linguistischen Konzepte und Methoden werden zuerst eingefügt. Die Notwendigkeit der Einführung „semantischer Einheit“ wird aber nicht deutlich erläutert. Um ein Modell für Abbildungen zu erstellen, werden aus dem OpenSubtitles-Korpus drei Trainingsmengen als Trainingsdaten erzeugt: Sätze, Einheiten und Lemmata. Schließlich werden die Performanzen des Modells tabellarisch vorgestellt, und die Verbesserungsmöglichkeiten vorgeschlagen.

Das spezielle Problem der Polysemie von Adpositionen wird im Kapitel 4 behandelt. Auch in diesem Kapitel werden zuerst die gängige Gruppierung und Klassifizierung von Adpositionen in der Linguistik vorgestellt. Acht deutsche Adpositionen werden für die Arbeit gewählt und ihre Verteilung im Korpus wird angegeben. Die

semantischen Klassen zur Annotation von Adpositionen werden in §4.3 präsentiert. Dann wird durch die Untersuchung des Korpus die Mehrdeutigkeit der Adpositionen in semantischen Einheiten festgestellt. Drei Modelle werden für Deutsch, Polnisch und Deutsch-Polnisch entsprechend trainiert. Die Präzisionen aller Modelle zeigen, dass die Mehrdeutigkeit von Adpositionen durch Hinzunahme einer weiteren Sprache reduziert werden kann. Eine Diskussion über die Verbesserung der Genauigkeit ist am Ende des Kapitels gegeben.

In der vorgelegten Bachelorarbeit hat sich Herr Tim Nieradzik mit einem komplexen Thema in der Computer-Linguistik beschäftigt. Er hat sich gut und weitergehend in das Themengebiet eingearbeitet, zahlreiche Literaturen wurden studiert und in der Arbeit eingeführt. Die Präsentation der eigenen Arbeit ist dagegen relativ knapp, insbesondere wurden die Ergebnisse nicht ausreichend erläutert. Wegen der allgemeinen Komplexität automatischer Generierung einer Wissensbasis sind die erzielten Resultate stark theoriebezogen, deren praktische Anwendungen wurden nicht explizit diskutiert.

Daher bewerte ich die Arbeit mit der Note: Gut (2,0)

Dr. Hui Shi