Universität Bremen

✉ Universität Bremen · **FB3 Informatik** / **TZI** · Postfach 33 04 40 · 28334 Bremen

FB3 Informatik / TZI

Dr. habil.
Hagen Langer
Wissenschaftlicher Mitarbeiter AG KI

Am Fallturm 1
28359 Bremen

Telefon (0421) 218 - 64011
Fax (0421) 218 - 64047
eMail hlanger@tzi.de
www www.tzi.de

Datum: 10.2.2015

Gutachten zur Bachelor-Arbeit
”Automatische Generierung einer Wissensbasis mithilfe mehrsprachiger Korpora”
vorgelegt von Tim Nieradzik

Die vorliegende Bachelor-Arbeit widmet sich dem Problem der Informationsextraktion aus natürlichsprachlichen Texten, und zwar speziell der Extraktion von Informationen aus mehrsprachigen Korpora. Dem Titel zufolge sollen diese Informationen bei der automatischen Generierung einer Wissensbasis genutzt werden.

Nach einer äußerst knappen Einleitung (Kapitel 1), in der die Fragestellung umrissen und ein Überblick über die Struktur der Arbeit gegeben wird, widmet sich das zweite Kapitel (”Annotieren von Korpora”) zunächst der Definition einiger Basisbegriffe (u.a. ”Tagset” und ”Morphem”) und beschreibt dann einige grundlegende Methoden wie Kompositazerlegung und Tagging; ferner werden Textkorpora thematisiert. Das Kapitel endet mit einem Abschnitt (”Realisierung”), in dem die gewählten Ressourcen (Korpus, Tagger, morphologische Analyse) sowie in einigen wenigen Sätzen (2.5.5) die Resultate der Arbeit beschrieben werden.

Kap. 3 (”Einheiten und deren Abbildung”) diskutiert zunächst unterschiedliche Ansätze der Syntax (u.a. Konstituentenstrukturen und Dependenzgrammatiken) und der Semantik (Rollen und frames) sowie das Alignment (in der Arbeit mit ”Abbildung” übersetzt) von parallelen Korpora.

Im vierten Kapitel (”Semantik von Adpositionen”) werden zunächst allgemeinere Eigenschaften von Adpositionen (vor allem: Präpositionen) diskutiert und dann ein Klassifikationsschema entwickelt, das sich an den in Literatur verwendeten Merkmalen (temporal, instrumental usw.) orientiert. Annotationsprobleme werden anhand zahlreicher Beispiele illustriert. Auf ca. 2 Seiten wird dann ein Experiment mit einem Entscheidungsbaumlerner beschrieben. Das Kapitel schließt mit einem Ausblick.

Kap. 5 gibt schließlich eine äußerst knappe Zusammenfassung.

Die Auswahl der Textgrundlage (Filmuntertitel) wird nur ansatzweise begründet und

wäre aber gerade wegen der bei ”ersten Analysen” (S. 13) festgestellten Mängel (OCR-Fehler) besser zu motivieren gewesen, weil derartige Mängel bei den anderen anführten Textkorpora (z.B. Wikipedia) eher nicht zu erwarten sind. Auch inwiefern diese Textgrundlage für die Erzeugung von Wissenbasen geeignet ist, hätte zumindest knapp diskutiert werden müssen. Auch die Auswahlkriterien für die Merkmale (S. 47) bleiben leider unklar. Die Verwendung von Entscheidungsbäumen (S. 49) wird lediglich dadurch motiviert, dass alle Merkmale nominal sind.

Die Arbeit ist überstrukturiert, so bestehen z.B. das erste und das letzte ”Kapitel” jeweils aus weniger als 20 Zeilen und auf S. 33 gibt es 5 Überschriften mit Abschnitten, die jeweils nur aus wenigen Zeilen bestehen, Abschnitt 2.5.3 besteht lediglich aus einem Satz, ab S. 42 gibt es Unterabschnitte, die nicht einmal einen einzigen vollständigen Satz enthalten usw. Auch wegen dieser übertriebenen Gliederung der Arbeit in kleinste Textstücke ist es gelegentlich schwierig einen roten Faden zu erkennen. Hier wäre eine Orientierung an Standardgliederungen wissenschaftlicher Arbeiten (z.B.: Fragestellung, Motivation, Stand der Forschung, Eigener Ansatz, Evaluierung, Ausblick) sicher empfehlenswert gewesen.

Das im Titel der Arbeit formulierte Thema ”Automatische Generierung einer Wissensbasis” wird in der Arbeit nur am Rande gestreift. Hier wäre zumindest ein perspektivischer Bezug – etwa im ”Ausblick” – herzustellen gewesen (z.B. ob und wie die Ergebnisse der Arbeit als Beitrag oder Vorarbeiten zur Lösung dieses Problems angesehen werden können).

Gelegentlich wird die linguistische Terminologie nicht ganz korrekt verwendet (”Präpositionsphrase”, S. 39, ”Lemmas” statt ”Lemmata”, ”Flektierung” statt ”Flexion”).

Herr Nieradzik hat eine komplexe und vielschichtige Aufgabenstellung bearbeitet, die neben informatischen Herausforderungen zusätzliche Kompetenzen im Bereich Linguistik und Sprachtechnologie erfordert. Dabei ist es ihm gelungen, sich selbstständig in die Materie einzuarbeiten und ein Teilproblem aus dem Themenbereich Informationsextraktion aus mehrsprachigen Korpora zu bearbeiten. Besondere Stärken der Arbeit sind die z.T. sehr detaillierten linguistischen Analysen und guten Einzelbeobachtungen (z.B. in 4.4.1) sowie die eigenständige Einarbeitung in ein breites und hochkomplexes Themenfeld. Leider fehlt der Arbeit eine klare Linie, die ausgehend von einer gut definierten Fragestellung zu einer kohärenten, zielgerichteten und überzeugend begründeten Auswahl von Daten und Methoden führt. Viele Entscheidungen bleiben letztlich unmotiviert und wirken ad hoc. Der Bezug zu Anwendungen bei der automatischen Generierung von Wissensbasen wird nicht in überzeugender Form hergestellt.

Insgesamt beurteile ich diese Arbeit mit der Note 2.0 (gut).

H. Langer