Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MIN-Fakultät
Fachbereich Informatik
Arbeitsbereich Natürlich-sprachliche Systeme (NatS)

Umwandlung der Hamburg Dependency Treebank in das Universal Dependency Format

Abschlussarbeit (Bachelor)

Kurzfassung: Am Arbeitsbereich NatS wurde eine Treebank annotiert[2] (d.h. Sätze von heise.de wurden mit syntaktischer Information angereichert). Die Treebank umfasst über 200.000 manuell annotierte Sätze. Universal Dependencies[3] sind ein u.a. mit Unterstützung von Google entworfenes Schema für syntaktische Annotationen, welches nicht sprachabhängig ist. Wie kann man die HDT automatisch nach Universal Dependencies konvertieren und welche Informationen gehen verloren?

Um eine Baumbank zu erstellen, wird meist ein mehr oder weniger langes Annotationsschema aufgestellt, welches genau erklärt, wie Sätze anotiert werden soll. Für die HDT ist dies Foth [1]. Zeman u. a. [4] haben bereits Baumbanken für mehrere Sprachen automatisiert nach Universal Dependencies übersetzt. Bei so einer Übersetzung geht immer Information verloren, wir wissen aber nicht, *welche* Informationen verloren gehen und ob diese wenigstens manuell nachgepflegt werden könnten.

Ziel der Abschlussarbeit soll es sein, ein Programm zu schreiben, welches die Hamburg Dependency Treebank nach Universal Dependencies umwandelt. Dabei soll insbesondere untersucht werden,

- Welche Unterschiede zwischen den Annotationsarten bestehen.
- Welche Konstrukte sich nicht übersetzen lassen (und somit Information verloren geht)
- Welche Konstrukte sich nicht *automatisiert* übersetzen lassen.

Es ist *nicht* Ziel der Arbeit, tatsächlich eine komplette Übersetzung zu erzeugen, insbesondere, wenn dafür manuelle Arbeit notwendig wäre.

Du brauchst Kenntnisse in einer Programmiersprache deiner Wahl (Perl, wenn du existierende Software aus Prag erweitern möchtest) und solltest Spaß an Algorithmen / formaler Informatik haben (die automatische Übersetzung ist effektiv ein Transduktor, also eine Grammatik, die eine Annotation in eine andere überführt). Ein Interesse an natürlicher Sprache ist ebenfalls sinnvoll, Vorkenntnisse aber nicht nötig.

Die Abschlussarbeit kann auf Deutsch oder Englisch erarbeitet und verfasst werden. Teile der relevanten Literatur sind nur auf Englisch verfügbar. Bei jeder Abschlussarbeit streben wir die Veröffentlichung der Forschungsergebnisse auf einer internationalen oder nationalen Konferenz an.

Literatur

[1] Kilian A. Foth. *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. ger. 2006. URL: http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/204/.

[2] Kilian A. Foth u. a. „Because Size Does Matter: The Hamburg Dependency Treebank". In: *Proceedings of the Language Resources and Evaluation Conference 2014*. LREC. European Language Resources Association (ELRA), 2014. URL: http://http://nats-www.informatik.uni-hamburg.de/HDT/.

[3] Joakim Nivre. „Towards a Universal Grammar for Natural Language Processing". English. In: *Computational Linguistics and Intelligent Text Processing*. Hrsg. von Alexander Gelbukh. Bd. 9041. Lecture Notes in Computer Science. Springer International Publishing, 2015, S. 3–16. ISBN: 978-3-319-18110-3. DOI: 10.1007/978-3-319-18111-0_1. URL: http://dx.doi.org/10.1007/978-3-319-18111-0_1.

[4] Daniel Zeman u. a. „HamleDT: To Parse or Not to Parse?" Englisch. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Hrsg. von Nicoletta Calzolari (Conference Chair) u. a. Istanbul, Turkey: European Language Resources Association (ELRA), Mai 2012. ISBN: 978-2-9517408-7-7.

Kontakt

Arne Köhn (*koehn@inf...*), Prof. Wolfgang Menzel

URL dieses Dokuments: