DGfS-CL

Thomas Proisl, Peter Uhrig
Interdisziplinäres Zentrum für Lexikografie, Valenz- und Kollokationsforschung, FAU Erlangen-Nürnberg

Korpora mit dem *Treebank.info*-Projekt syntaktisch parsen und abfragen

Das Erlanger *Treebank.info-Projekt* hat es sich zum Ziel gesetzt, automatisch dependenz-annotierte Korpora auch für computerlinguistische Laien nutzbar zu machen. Zu diesem Zweck wurden ein hochskalierbarer Parsing-Dienst sowie ein Backend für die effiziente Beantwortung von Abfragen über ein intuitives grafisches Web-Interface verfügbar gemacht. Die Parsing-Infrastuktur kann durch die Verwendung einer leistungsstarken Dokumenten-datenbank (MongoDB) in Verbindung mit asynchroner Kommunikation per Messaging Middleware (HornetQ) und der Nutzung verschiedener High-Performance-Computing-Systeme im Erlanger Rechenzentrum mit über 1000 CPU-Kernen parallel parsen, so dass auch große Datenmengen selbst mit vergleichsweise langsamen Parsern in wenigen Stunden verarbeitet werden können. Dabei ist für die plattformunabhängigen Parsing-Clients kein Installationsaufwand nötig, so dass z.B. auch ein nicht ausgelasteter Computerraum in der vorlesungsfreien Zeit zum Parsen verwendet werden könnte.
Zur effizienten Abfrage von Dependenzstrukturen wurde eine Lösung namens CWB-treebank (Proisl/Uhrig 2012) auf Basis der IMS Open Corpus Workbench (Christ 1994; Christ/Schulze 1995) implementiert, die viele komplexe Anfragen in einer für ein Web-Interface angemessenen Geschwindigkeit beantworten kann. Darüber hinaus besteht die Möglichkeit, sich Kollokationen auf Basis von beliebig definierbaren Gruppen von Dependenzrelationen vorberechnen zu lassen, die dann verzögerungsfrei angezeigt werden können (detaillierter beschrieben und evaluiert in Uhrig/Proisl (im Erscheinen, 2012)).
Die Präsentation in der Postersession der Sektion Computerlinguistik verfolgt zwei Ziele. Einerseits soll die Architektur des Systems den computerlinguistischen Kollegen vorgestellt werden, andererseits soll eine Softwaredemonstration dazu dienen, theoretischen und angewandten Sprachwissenschaftlern zu demonstrieren, wie sie eigene Korpora hochladen und parsen lassen können, um diese dann per Web-Browser abzufragen.

Literatur:

Christ, Oliver (1994): „A modular and flexible architecture for an integrated corpus query system.“ In: *Proceedings of COMPLEX’94: 3rd Conference on Computational Lexicography and Text Research*. Budapest, 23–32.

Christ, Oliver, Bruno M. Schulze (1995): „Ein flexibles und modulares Anfragesystem für Textcorpora.“ In: *Tagungsberichte des Arbeitstreffens Lexikon + Text*. Tübingen: Niemeyer.

Proisl, Thomas, Peter Uhrig (2012): „Efficient Dependency Graph Matching with the IMS Open Corpus Workbench.“ In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Hrsg.): *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul: European Language Resources Association (ELRA), 2750–2756.

Uhrig, Peter, Thomas Proisl (im Erscheinen, 2012): „Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates.“ In: *Lexicographica* 28.