Verbesserung der Kollokationsextraktion durch Dependenzenannotation von Korpora

Thomas Proisl & Peter Uhrig
Interdisziplinäres Zentrum für Lexikografie, Valenz- und Kollokationsforschung
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bismarckstr. 1, 91054 Erlangen
Thomas.Proisl@linguistik.uni-erlangen.de
Peter.Uhrig@angl.phil.uni-erlangen.de

Die Dependenzgrammatik hat in den vergangenen Jahren verstärktes Interesse in der maschinellen Sprachverarbeitung erfahren, sodass es heute sowohl native Dependenzparser als auch Programme zur Konvertierung von Phrasenstrukturen in Dependenzstrukturen gibt. Da eine Dependenzanalyse die syntaktisch-semantischen Relationen zwischen den Wortformen eines Satzes explizit macht, bietet sie ideale Voraussetzungen für die automatische Extraktion von Kollokationen und anderen Mehrworteinheiten aus Textkorpora (vgl. Weller/Heid 2010). Durch die Dependenzannotation kann die Untersuchung zielgerichtet auf diejenigen Kookkurenzpartner eingeschränkt werden, die zueinander in Beziehung stehen, wodurch Formen, die nur zufällig im unmittelbaren Kontext stehen, herausgefiltert werden.

In diesem Vortrag soll die Extraktion von Kollokationen auf Basis von Dependenzrelationen mit der herkömmlichen fensterbasierten Kollokationsextraktion verglichen werden. Als Datengrundlage dient dabei die geparste Version des British National Corpus des Treebank.info-Projekts. Die qualitative Evaluation der gefundenen Kollokationskandidaten (vgl. Evert/Krenn 2001) wird dabei anhand des Oxford Collocations Dictionary for Students of English durchgeführt.

Literatur

Evert, Stefan/Brigitte Krenn (2001): Methods for the Qualitative Evaluation of Lexical Association Measures. In: ACL'01, S. 188-195.

Oxford Collocations Dictionary for Students of English (2002), (editor Diana Lea) Oxford u.a.: OUP.

Uhrig, Peter/Thomas Proisl (2011): Treebank.info. http://treebank.info.

Weller, Marion/Ulrich Heid (2010): Extraction of German Multiword Expressions from Parsed Corpora Using Context Features. In: LREC'10, S. 3195-3201.