Berlin-Brandenburgische Akademie
der Wissenschaften (BBAW),
Berlin

Technischer Bericht

Nr. 2016/1 (Meilenstein 4b)

KobRA-Integration in die Rechercheumgebung am Zentrum Sprache der BBAW

BMBF-Verbundprojekt:

Korpus-basierte linguistische Recherche und Analyse mithilfe von Data-Mining (KobRA)

Förderkennzeichen: 01UG1245B
Projektlaufzeit: 01.09.2012 bis 31.08.2015
Bearbeiter/innen: Alexander Geyken, Bryan Jurish, Kay-Michael Würzner

Berlin, 24.2.2016

Das diesem Bericht zugrunde liegende Vorhaben wird mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter den Förderkennzeichen 01UG1245A-D gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

KobRA-Integration in die Rechercheumgebung am Zentrum Sprache der BBAW

1. Linguistische Recherche-Umgebung des Zentrums Sprache der BBAW

Das Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften verfügt über verschiedene Korpora geschriebener Sprache (Diachrone Referenzkorpora, gegenwartssprachliche Zeitungskorpora, Blogkorpora, etc.) mit einem Gesamtumfang von ca. 8 Milliarden laufenden Wörtern. Diese Korpora stehen projektintern für die (computer-)lexikographischen Aufgaben zur Verfügung. Eine Teilmenge der Korpora (etwa 1,5 Milliarden laufender Textwörter) stehen öffentlich auf der Webplattform (www.dwds.de) zur Verfügung. Die Korpora wurden mit sprachtechnologischen Werkzeugen linguistisch annotiert. Für jedes Wort werden das Grundwort (Lemma) und die Wortart angegeben (Geyken 2007, Klein & Geyken 2010).

Alle Korpora sind mit der Suchmaschine DDC (Sokirko, 2003) indiziert und abfragbar. Die öffentlich verfügbaren Korpora sind in die CLARIN-Infrastruktur integriert und über die "Föderierte Content Suche" (FCS) abfragbar. Mit der DDC Abfragesprache können korpusübergreifend Wörter, Wortverbindungen und linguistische Strukturen abgefragt werden. Die Ergebnisse werden als "Keyword in Context"-Konkordanzen angezeigt.

Auf dieser Infrastruktur aufbauend wurden am Zentrum Sprache zwei Werkzeuge entwickelt und implementiert, die vor allem die Belange der korpusbasierten Lexikographie, aber auch z.B. das Sprachlernen unterstützen.

(i) Das "DWDS-Wortprofil" extrahiert zu einem Stichwort (der Basis) die statistisch signifikanten Kookurrenzen (also gemeinsam in einem Satz vorkommende Wortpaare bzw. -tripel). Darüber hinaus werden diese nach der Art der syntaktischen Beziehung, in der die Wortpaare bzw. -tripel stehen, klassifiziert und so gruppiert angezeigt. Für jede Wortverbindung ist die Anzeige der Belege für diese Verbindung vorgesehen. Verschiedene Stichwörter können anhand der Gemeinsamkeiten und Unterschiede der Wortverbindungen verglichen werden (Geyken et al. 2009; Lemnitzer & Geyken 2015).

(ii) Der "Gute-Belege-Extraktor" (GBE) bewertet alle Belege zu einem Stichwort mit einer "Gütefunktion" und ordnet sie in absteigender Reihenfolge an (Didakowski et al. 2012). Die Gütekriterien sind dabei als Regeln und Präferenzen formuliert. Der Zweck des Gute-Belege-Extraktor besteht darin, den den Lexikographinnnen und Lexikographen die Arbeit zu erleich-

tern, indem für das Wörterbuch ungeeignete Belege einen niedrigen "Gütewert" erhalten sollen, "gute" Belege hingegen einen hohen. Bei Belegmengen, die häufig im 5 bis 6-stelligen Bereich liegen, kann der Gute-Belege-Extraktor eine große Arbeitserleichterung bringen (Lemnitzer et al. 2015). Die besten Belege zu jedem Stichwort werden auf der Webseite des Projekts (www.dwds.de) für alle Stichwörter angezeigt, die bislang noch nicht semantisch bearbeitet wurden.

Sowohl das DWDS-Wortprofil als auch der Gute-Beispiele-Extraktor werden auch als Webservices zur Integration in andere Umgebungen bereitgestellt.

Das Korpusverwaltungssystem "D*" des Zentrums Sprache wurde entwickelt, um einen einheitlichen Zugang zu den verschiedenen hier angesiedelten Korpora zu ermöglichen. Im Rahmen des KobRA Projekts wurde das D* System um einen standardisierten RESTful API (Fielding 2000) erweitert, um automatisierte Korpusabfragen von den in Dortmund entwickelten Softwarewerkzeugen zu ermöglichen. Um die Kommunikation zwischen der Suchmaschine und den KobRA-Werkzeugen zu optimieren wurde die Korpussuchmaschine DDC (Sokirko 2003, Jurish et al. 2014) weiterentwickelt, um jeden Suchterm mit einem benutzerspezifierten Trefferindex (Subskript oder sog. "Match-ID") zu versehen. Desweiteren wurde für selegierte Korpora einen Zugang über die in Tübingen enwickelte WebLicht Umgebung geschaffen.

2. Integration von Verfahren des maschinellen Lernens in den lexikographischen Arbeitsprozess des DWDS

Die Basis zeitgemäßer lexikographischer Arbeit ist die Rückbindung von Verwendungsbeispielen und Wortbedeutungen an Korpusbelege. Die zunehmende Verfügbarkeit digitaler Korpora (s.o.) erhöht auf der einen Seite die Menge an Textmaterial aus dem der Lexikograph schöpfen kann, verlangsamt aber auf der anderen Seite auch den Auswahlprozess, da mehr und mehr Textstellen einer genauen Überprüfung unterzogen werden müssen. Eine automatische Filterung von Korpusbelegen bzw. deren Vorauswahl anhand wohldefinierter Kriterien ist daher unabdinglich für die korpusbasierte lexikographische Arbeit. Für die Belegauswahl bei der Erstellung neuer und die Aktualisierung vorhandener Artikel kommt im Digitalen Wörterbuch der deutschen Sprache (DWDS) ein regelbasiertes Belegextraktionsverfahren (Didakowski et al. 2012) zum Einsatz, das im Wesentlichen auf dem bahnbrechenden und mittlerweile in vielen lexikographischen Projekten eingesetzten Ansatz von Kilgarriff et al. (2008) basiert. Trotz des unbestrittenen positiven Einflusses der automatischen Belegauswahl

hat sich in der täglichen lexikographischen Arbeit gezeigt, dass die ausgewählte Menge von 15--20 Belegen häufig noch zu viele unerwünschte Treffer enthält. An dieser Stelle setzt die Integration der im KobRA-Projekt entwickelten Infrastruktur zum komfortablen Einsatz maschineller Lernverfahren in klassische lexikographische (und linguistische) Arbeitsprozesse ein: Anhand manuell bewerteter Korpusbelege sollen Eigenschaften gelernt werden, die einen guten von einem schlechten Beleg unterscheiden. Als zugrundeliegendes Modell kommen *Support Vector Machines* (SVMs, Joachims 1998) zum Einsatz. Formal kann die Klassifizierung wie folgt gefasst werden:
sign(<b, x> + b_0). Dabei handelt es sich um das Vorzeichen der Länge der Projektion der Belege als Merkmalsvektor x auf die Hyperebene gegeben durch den Normalenvektor b, die durch die SVM bestimmt wurde, plus einen Biasterm b_0.

Das Verfahren wurde in drei Varianten ausführlich in Lemnitzer et al. (2015) beschrieben. In der in der DWDS-Infrastruktur eingesetzen Variante werden die zu klassifizierenden Belege zunächst auf syntaktischer Ebene mit Hilfe des Stanford-Parsers stukturiert. Der Zugriff auf den Parser erfolgt dabei über das CLARIN-D-Werkzeug *Weblicht*. Abbildung 1 stellt das Zusammenspiel zwischen der KobRA-Infrastruktur, den DWDS-Werkzeugen und der CLARIN-D-Infrastruktur schematisch dar. Die Anwendung im lexikographischen Arbeitsprozess des DWDS bestätigt die grundsätzlich positive Evaluation (vgl. Lemnitzer et al. 2015): der maschinelle Lerner filtert aus der vom Gute-Beispiele-Extraktor produzierten Belegmenge einen höheren Anteil an “schlechten” Beispielen als an “guten” Belegen. Die Qualität der Belege wird somit durch den maschinellen Lerner weiter erhöht und verbessert somit weiter den lexikographischen Prozess der Belegauswahl (GBE, s. oben).

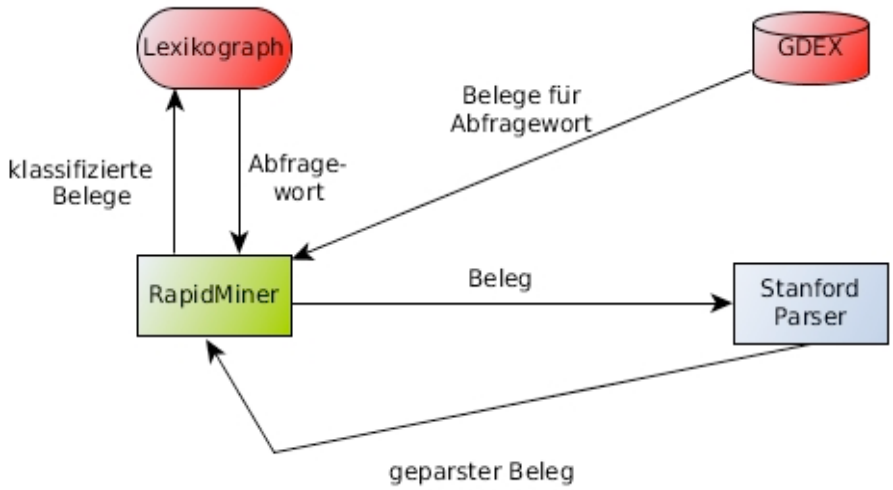

Abbildung 1: Workflow zur Klassifizierung von Korpusbelegen.

Als weiteres mögliches Einsatzgebiet für Techniken des maschinellen Lernens innerhalb des lexikographischen Arbeitsprozesses wurde im Projektverlauf die automatische Zuordnung von Korpusbelegen für polyseme Wörter zu deren im Wörterbuch beschriebenen Bedeutungen untersucht. Eine zuverlässig funktionierende Lösung dieses Spezialfalls der sog. *Word Sense Disambiguation* (Weaver, 1949) könnte die Wörterbuchartikelredaktion zusätzlich vereinfachen und beschleunigen. Als Ausgangspunkt für die Untersuchung diente ein Ansatz von Lesk (1986), der im Wesentlichen anhand der Übereinstimmung des Wortmaterials in Bedeutungsbeschreibung und Korpusbeleg die Zuordnung vornimmt. Dazu wurden im Projekt zwei Erweiterungen vorgeschlagen: Zunächst werden neben der direkten Übereinstimmung des Wortmaterials in Definition und Beleg auch Übereinstimmungen in den Wortprofilen (s.o.) der jeweils enthaltenen Inhaltswörter herangezogen, um die Größe der Schnittmengen zu erweitern. In einem weiteren Schritt wurde auf Basis manuell erzeugter Zuordnungsbeispiele ein *Maximum Entropy Classifier* (Nigam et al. 1999) trainiert, der die im Wortprofil vorhandenen syntaktischen Relationen gewichtet und so ihren Einfluss auf die Bedeutungszuordnung idealiter optimal ausbalanciert. Details zu diesem Vorgehen finden sich in Geyken et al. (2015). Für die Klassifizierung wurde wiederum auf die KobRA-Infrastruktur zurückgegriffen. Der diesbzgl. Workflow ist in Abbildung 2 dargestellt.

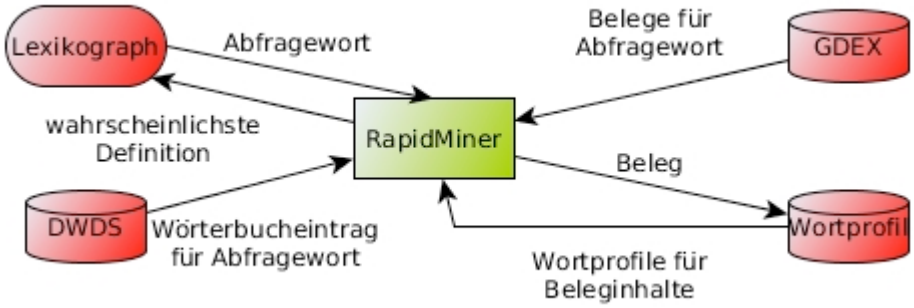

Abbildung 2: Workflow zur Bedeutungsdisambiguierung polysemer Wörter

3. Zitierte Literatur

Didakowski, J., Geyken, A. & Lemnitzer, L. (2012). Automatic example sentence extraction for a contemporary German dictionary. In: R. Vatvedt Fjeld & J. M. Torjusen (Hrsg.), *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012* (S. 343-349). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo. Abgerufen unter http://www.euralex.org/proceedings-toc/euralex_2012/

Fielding, R.T. (2000). Architectural Styles and the Design of Network-based Software Architechtures. PhD thesis, University of California, Irvine, 2000. Abgerufen unter https://www.ics.uci.edu/ ~fielding/pubs/dissertation/top.htm.

Geyken, A., Pölitz, C. & Bartz, T. (2015). A machine learning method based on word profiles for semi-automatic update of polysemous dictionary entries in legacy dictionari-

es. Kosem, I., Jakubíček, M., Kallas, J., Krek, S. (Hrsg.), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. (S. 304-314). Abgerufen unter https://elex.link/elex2015/proceedings/eLex_2015_19_Geyken+Politz+Bartz.pdf

Geyken A.; Didakowski, J.; Siebert, A. (2009). Generation of word profiles for large German corpora. In Y. Kawaguchi, M. Minegishi & J. Durand (Hrsg.), *Corpus Analysis and Variation in Linguistics* (S. 141-157). Tokio: Benjamins.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Hrsg.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects* (S. 23–41). London: Continuum Press.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In C. Nédellec & C. Rouveirol (Hrsg.), *Lecture Notes in Computer Science* (S. 137–142). Berlin: Springer.

Jurish, B.; Thomas, C. & Wiegand, F. (2014). Querying the deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner & C. Gurrin (Hrsg.), *Proceedings of the Workshop "Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities" (MindTheGap 2014)* (S. 25-30). Berlin, Germany, March 2014. Abgerufen unter: http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf.

Kilgarriff, A.; Husák, M.; McAdam, K.; Rundell, M. & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Elisenda Bernal & Janet DeCesaris (Hrsg.), *Proceedings of the XIII EURALEX International Congress* (S. 425-433). Barcelona: Universitat Pompeu Fabre.

Klein, W. & Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In R. H. Gouws u.a. (Hrsg.), *Lexicographica. Internationales Jahrbuch für Lexikographie* (Bd. 26, S. 79–96). Berlin: de Gruyter.

Lemnitzer, L. & Geyken, A. (2015). Semantic Modeling of Collocations for Lexicographic Purposes. *Journal of Cognitive Science*, 16 (3): 200-223.

Lemnitzer, L., Pölitz, C., Didakowski, J. & Geyken, A. (2015). Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German. In Kosem, I., Jakubíček, M., Kallas, J., Krek, S. (Hrsg.), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. (S. 21-31). Abgerufen unter https://elex.link/elex2015/proceedings/eLex_2015_02_Lemnitzer+etal.pdf

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In V. DeBuys (Hsrg.), *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*. (S. 24-26) New York, NY, USA. ACM.

Nigam, K., Lafferty, J. & McCallum, J. (1999). Using maximum entropy for text classification. In: *IJCAI-99: Workshop on Machine Learning for Information Filtering*, 61-67. Abgerufen unter http://www.kamalnigam.com/papers/maxent-ijcaiws99.pdf.

Sokirko, A. (2003). A technical overview of DWDS/Dialing Concordance. Vortrag im Rahmen der Konferenz Computational linguistics and intellectual technologies, Protvino, Russia. Abgerufen unter http://www.aot.ru/docs/OverviewOfConcordance.htm.

Weaver, W. (1949). Translation. In W.N Locke & A.D. Booth (Hrsg.). *Machine Translation of Languages: Fourteen Essays* (S. 15-23). Cambridge, MA: MIT Press.