Disambiguierung in Suchtrefferlisten aus großen Textkorpora: Anwendungsfelder und Perspektiven

Thomas Bartz[1], Alexander Geyken[3], Christian Pölitz[2], Achim Saupe[4], Angelika Storrer[1]
Technische Universität Dortmund, Institut für deutsche Sprache und Literatur[1] / Fakultät Informatik[2]
Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), Zentrum Sprache[3]
Zentrum für Zeithistorische Forschung Potsdam (ZZF)[4]

1. Zielsetzung und Projekthintergrund

Digitale Textkorpora bieten in vielen geisteswissenschaftlichen Arbeitsbereichen neuartige Möglichkeiten, Forschungsfragen an authentischen Sprachverwendungen zu untersuchen. Infrastrukturprojekte wie CLARIN bieten flexible Werkzeuge zur Datengewinnung und zur quantitativen Analyse an, mit denen große, linguistisch strukturierte Textkorpora ausgewertet werden können. Allerdings müssen die automatisch gewonnenen Daten oft noch manuell nachbearbeitet werden. Dies ist insbesondere der Fall, wenn nicht Wortformen, sondern sprachliche Zeichen – also Verbindungen von Form und Inhalt – quantitativ ausgewertet werden sollen, denn homonyme und polyseme Textwörter sind in aktuell verfügbaren Korpora nicht disambiguiert. Wenn die Unterscheidung von Homonymen und polysemen Lesarten für eine Forschungsfrage relevant ist, müssen die Daten bislang manuell disambiguiert werden. Der damit verbundene Aufwand ist oft erheblich; unter den zeitlichen Restriktionen eines Forschungsprojekts, einer Dissertation, einer studentischen Abschlussarbeit etc. können so bestimmte Fragestellungen gar nicht bearbeitet werden.

Im Verbundprojekt „Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining“ (KobRA, http://www.kobra.tu-dortmund.de) arbeiten germanistische Linguistik, Informatik, Sprachtechnologie und Sprachressourcenanbieter gemeinsam daran, den Aufwand der manuellen Nachbearbeitung zu senken und damit die Möglichkeiten der korpusbasierten Recherche und Analyse zu verbessern.[1] Dazu werden Machine-Learning- und Data-Mining-Verfahren des Informatikpartners für Aufgaben aus aktuellen Forschungsvorhaben der Linguistik angepasst und in Fallstudien erprobt. Die beteiligten Sprachtechnologie- und Sprachressourcenpartner stellen dazu unterschiedlich strukturierte große digitale Textkorpora bereit (z.B. wortartenannotierte Korpora, Baumbanken etc.) und integrieren die entwickelten Verfahren in die vorhandene Infrastruktur. Die Fallstudien beziehen sich auf drei Anwendungsfelder: Korpusbasierte Lexikographie, diachronische Sprachforschung und Varietätenlinguistik. Bei den daraus abgeleiteten Aufgabenstellungen handelt es sich um Routineaufgaben bei der Arbeit mit großen Textkorpora (Filtern, Klassifizieren, Disambiguieren, Visualisieren), die sich in verschiedenen geisteswissenschaftlichen Arbeitsbereichen in ähnlicher Form stellen.

Im Vortrag erläutern wir zunächst ausgehend von einem konkreten Anwendungsszenario aus der korpusbasierten Lexikographie Herausforderungen, die bei der Arbeit mit großen Textkorpora entstehen, und leiten daraus Anforderungen an mögliche automatische Verfahren ab. Im Anschluss stellen wir erste erprobte Verfahren, verwendete Korpora und bisher erzielte Ergebnisse vor. In einem dritten Schritt zeigen wir schließlich perspektivisch den analytischen Mehrwert der Verfahren auch für Aufgabenstellungen der historischen Forschung auf.

[1] Das Verbundprojekt wird vom Bundesministerium für Bildung und Forschung (BMBF) seit Herbst 2012 im Rahmen des Programms „eHumanities“ gefördert. Beteiligt sind die folgenden Institutionen und Projektleiter: TU Dortmund (Germanistik: Angelika Storrer, Informatik: Katharina Morik), Berlin-Brandenburgische Akademie der Wissenschaften (Alexander Geyken), Eberhard-Karls-Universität Tübingen (Erhard Hinrichs), Institut für deutsche Sprache, Mannheim (Marc Kupietz/Andreas Witt).

2. Fallstudie im Anwendungsfeld korpusbasierte Lexikographie

2.1 Anwendungsszenario

Ein wichtiges Einsatzgebiet für digitale Textkorpora ist seit langem die Sprachlexikographie (vgl. Engelberg/Lemnitzer 2009). In einem digitalen Referenzkorpus wie dem DWDS-Kernkorpus (vgl. Geyken 2007), das im Hinblick auf die Verteilung der enthaltenen Textbestände auf die Textsortenbereiche Belletristik, Gebrauchsliteratur, Wissenschaft und journalistische Prosa sowie auf die Dekaden des 20. Jahrhunderts ausgewogen ist, können Lexikographen zu einem Suchwort automatisch Daten zur Frequenzentwicklung über das 20. Jahrhundert hinweg gewinnen und die Gebräuchlichkeit des Wortes in verschiedenen Textsortenbereichen vergleichen. Wenn man allerdings Aussagen zur Textsortenspezifik und zur Bedeutungsentwicklung einer speziellen Wortbedeutung treffen möchte, müssen die vom System ausgegebenen Belege bei polysemen oder homonymen Lexemen manuell disambiguiert werden. Wenn sich die Anzahl der Treffer zu einem Suchwort in überschaubaren Grenzen hält, wie im Falle des in Storrer (2011) diskutierten Beispielworts *Ampel*, ist eine solche Disambiguierung noch mit vertretbarem Zeitraufwand möglich. Bei dem in 2.3. untersuchten Beispielwort *Leiter* resultiert die Suche im DWDS-Kernkorpus bereits in einer Liste mit 6895 Belegen, die in nicht vorhersehbaren Anteilen Belege für die Homonyme *der Leiter* und *die Leiter* sowie für speziellere Lesarten (z.B. *Leiter* i.S. von *Energieleiter*, *Trittleiter*, *Tonleiter*) enthält. Für viele Lexeme sind die Belegzahlen noch höher; eine manuelle Disambiguierung ist in solchen Fällen zeitlich extrem aufwändig. In unserer unten beschriebenen Fallstudie suchen wir deshalb nach Verfahren zur automatischen Disambiguierung von Suchwörtern in Belegen, wie sie von Korpusrecherchesystemen ausgegeben werden. Die Verfahren sollen das Arbeiten mit Korpora in der Lexikographie vereinfachen und verbessern. Weiterhin sollen auf ihrer Basis statistische Analyse- und Visualisierungswerkzeuge für Korpusdaten (z.B. Kookkurrenzanalysen, Wortverlaufsdiagramme), die bislang noch überwiegend formbasiert arbeiten, um Komponenten zur semantischen Disambiguierung angereichert werden. Zudem sollen Verfahren entwickelt werden, die auch für Suchwörter, die als monosem gelten, ungewöhnliche und/oder neuartige Verwendungen zu Tage fördern.

2.2 Verwandte Arbeiten

Das vorgestellte Anwendungsszenario liegt in Reichweite der Forschung zur automatischen Disambiguierung von Wortbedeutungen (Word-Sense-Disambiguation, WSD), der sich zahlreiche Arbeiten widmen, die hier nicht in ihrer Breite dargestellt werden können (für einen Überblick vgl. Agirre et al. 2007; eine umfangreiche Vergleichsstudie zu aktuellen Verfahren hat Navigli 2009 veröffentlicht). Sie scheinen in unserem Fall auch nicht zielführend zu sein, weil sie i.d.R. Wörter nach vorgegebenen Lesarten disambiguieren. Dadurch wäre aber die Möglichkeit, im Korpus potenziell enthaltene unerwartete Lesarten zu entdecken, von vornherein ausgeschlossen.

Wir folgen in unserer Studie daher einem Ansatz, bei dem die Lesarten induktiv aus den Korpusdaten ermittelt werden. Für diese Word-Sense-Induction (WSI) liegen eine Reihe erfolgreicher Ansätze vor, die im Wesentlichen auf Clustering-Verfahren basieren (für einen Überblick vgl. Brody/Lapata 2009). Brody und Lapata (2009) konnten zeigen, dass sich mithilfe der Latent-Dirichlet-Allocation (LDA, vgl. Blei et al. 2003) tendenziell die besten Ergebnisse erzielen lassen. LDA basiert auf der Annahme, dass dasselbe Wort in verschiedenen Lesarten verwendet wird, wenn es in unterschiedlichen Kontexten vorkommt. Dazu werden um alle Vorkommen eines zu behandelnden Wortes Kontextfenster in einer bestimmten Größe gelegt und mithilfe von Wort- und Kookkurrenzstatistiken Verteilungen von Kontextwörtern, sogenannte „Topics“, ermittelt, die als Lesarten aufgefasst werden können. Für jedes einzelne Kontextfenster lässt sich daraufhin die Wahrscheinlichkeit berechnen, mit der es einem bestimmten Topic bzw. ein Vorkommen des zu behandelnden Wortes einer bestimmten Lesart zugeordnet werden kann. Dabei wird angenommen, dass die Wahrscheinlichkeit für die Zuordnung zu den

Topics einer Dirichletverteilung folgt. Rohrdantz et al. (2011) zeigten den Nutzen des Verfahrens als Grundlage für Visualisierungen zur Bedeutungsentwicklung von Beispielwörtern aus einem Zeitungskorpus, die es erlauben, die Entstehung von Neubedeutungen und ihre Entwicklung über die Zeit zu rekonstruieren.

2.3 Erste Ergebnisse

Für die im Vortrag vorgestellte Fallstudie wird LDA an Sprachdaten aus dem DWDS-Kernkorpus des 20. Jahrhunderts (s. 1.) erprobt. Ergänzend zu Rohrdantz et al. (2011) können so weitere Erkenntnisse über den Nutzen des Verfahrens auch für deutsche Sprachdaten gewonnen werden, die zudem aus unterschiedlichen Textsortenbereichen stammen. Bislang wurde das Verfahren am Beispiel des Wortes *Leiter* evaluiert (6895 Belege aus dem DWDS-Kernkorpus, 2000 manuell disambiguierte Belege für die Evaluation; Verteilung der Lesarten s. Tabelle 1). Die Topics wurden zunächst ausschließlich auf Basis der Kontextwörter (Bag-of-Words) in einem noch verhältnismäßig großen Fenster im Umfang des jeweils ganzen das Wort *Leiter* enthaltenden Satzes ermittelt. Zur Evaluation des Verfahrens wurde die Reinheit der Topic-Cluster im Hinblick auf die manuell bestimmten Lesarten gemessen, als Maß diente NMI (Normalized Mutual Information, vgl. Manning et al. 2008: 357 f.).

Lesart	**Vorkommen**	
	absolut	**relativ**
Boss/Führungsperson	1665	0,833
Trittleiter	296	0,148
Energieleiter	29	0,015
Tonleiter	10	0,005
	2000	

Tabelle 1: Manuell ermittelte Lesarten

Wenngleich der angewandte Ansatz (NMI = 0,20) einem einfachen k-Means-Clustering (vgl. Lloyd 1957/1982; NMI = 0,16) bereits überlegen zu sein scheint, erfordert das beschriebene Anwendungsszenario (s. 2.1) eine weitere Verfeinerung des Verfahrens. Tabelle 2 veranschaulicht die Ergebnisse: Zum einen lassen sich die ermittelten Topics nur den zwei häufigsten manuell bestimmten Lesarten zuordnen (*Boss/Führungsperson* i.S.v. *politischer Leiter, DDR/Drittes Reich*: Topic 1/2; *Leiter einer Bildungsinstitution*: Topic 3; *musikalischer Leiter*: Topic 4; *Trittleiter*: Topic 5) – Belege für *Leiter* i.S.v. *Energie-* bzw. *Tonleiter* sind allerdings auch selten (s. Tabelle 1). Zum anderen sind die ermittelten charakteristischen Kontextwörter noch verhältnismäßig allgemein und ihre Salienz zu gering (z.B. im Vergleich zu syntaktischen Kookkurrenzstatistiken, vgl. Didakowski/Geyken 2013). Zur Verbesserung des Verfahrens werden deshalb gegenwärtig unterschiedliche Kontextfenster getestet sowie weitere Featureklassen (POS, Dependenzen) integriert. Eine ausführlichere Beschreibung und Auswertung der getesteten Ansätze würden wir sehr gerne im Paper bzw. im Vortrag präsentieren.

Topic 1	0,2329	**Topic 2**	0,2181	**Topic 3**	0,1821	**Topic 4**	0,1699	**Topic 5**	0,1969
DDR	0,0030	politisch	0,0040	Berlin	0,0027	Musik	0,0012	hinauf	0,0016
Abteilung	0,0027	Partei	0,0025	Prof.	0,0019	München	0,0011	Mann	0,0014
Regierung	0,0023	Korps	0,0019	Dr.	0,0012	New_York	0,0011	oben	0,0014
Minister	0,0016	Führer	0,0019	Hochschule	0,0009	Dirigent	0,0010	gehen	0,0012
ZK	0,0013	Arbeit	0,0017	Institut	0,0008	Oper	0,0008	Sprosse	0,0009
SED	0,0010	NSDAP	0,0010	Lehrer	0,0008	Komponist	0,0007	Wand	0,0008

Tabelle 2: Automatisch induzierte Topics und salienteste Kontextwörter (Auszug aus Top 50), jeweils mit Auftretenswahrscheinlichkeiten

3. Anwendungsmöglichkeiten des Verfahrens in anderen geisteswissenschaftlichen Disziplinen und Anwendungsbereichen: Das Beispiel „Historische Semantik des 20. Jahrhunderts“

Die Induktion von Lesarten auf Grundlage von Kontextwörtern stellt auch eine interessante Anwendungsmöglichkeit für die Geschichtswissenschaften dar. Am ZZF, mit dem die BBAW im Kontext des CLARIN-Projekts zusammenarbeitet, ist ein Projekt zur Historischen Semantik des 20. Jahrhunderts angesiedelt (vgl. Kollmeier/Hoffmann 2010, 2012; Kollmeier/Saupe im Erscheinen), welches sich mit dem Wandel von Begriffen und Topoi sowie den mit ihnen verbundenen Diskursen beschäftigt. Die quantitative Aufschlüsselung polysemer Begriffe ist auch hier ein bedeutsamer Gewinn: sie ermöglicht einen schnelleren Zugriff auf verschiedene Bedeutungen und Verwendungsweisen von historisch relevanten Begriffen und erleichtert damit die qualitative Auswertung von Textkorpora.

Im Rahmen der Kooperation zwischen der BBAW und dem ZZF soll das oben beschriebene Verfahren (s. 2.3) auf weitere Korpora angewandt werden. Insbesondere soll auf der Basis des digitalisierten Zeitungskorpus des DDR-Presseportals (Neues Deutschland, Berliner Zeitung, Neue Zeit, http://zefys.staatsbibliothek-berlin.de/ddr-presse/) die sprachliche Ambiguität von zentralen Begriffen untersucht werden. Das Beispiel des Begriffs *Einheit* verdeutlicht dies. Dieser stand in der DDR in verschwindendem Maße für die deutsche Einheit, seit 1946 dagegen vorrangig für die Einheit von SPD und KPD bzw. seit 1971 für die Einheit von Wirtschafts- und Sozialpolitik. Daneben tauchte der Begriff immer auch als Maßeinheit in der Produktionsberichterstattung auf. Das Disambiguierungsverfahren könnte dabei helfen, den historisch-semantischen Bedeutungswandel des Begriffs *Einheit* in der DDR empirisch auf quantitative Weise zu analysieren.

Literatur

Agirre, E. / Màrquez, L. / Wicentowski, R. (Hg.) (2007): Proceedings of the SemEval-2007. Prague, Czech Republic.

Blei, D. M. / Ng, A. Y. / Jordan, M. I. (2003): Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Brody, S. / Lapata, M. (2009): Bayesian word sense induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL ’09. Stroudsburg, PA, USA, 103–111.

Didakowski, J. / Geyken, A. 2013: From DWDS corpora to a German Word Profile – methodological problems and solutions. In: Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network „Internet Lexicography“. Mannheim: Institut für Deutsche Sprache (OPAL - Online publizierte Arbeiten zur Linguistik X/2012), 43–52.

Engelberg, S. / Lemnitzer, L. (2009): Einführung in die Lexikographie und Wörterbuchbenutzung. Tübingen: Stauffenburg.

Geyken, A. (2007): The DWDS corpus: a reference corpus for the German language of the 20th century. In: Fellbaum, C. (Hg.): Idioms and collocations. Corpus-based linguistic and lexicographic studies. London: Continuum, 23–40.

Kollmeier, K. / Saupe, A. (im Erscheinen): Ausgangspunkte einer Historischen Semantik des Politischen für das 20. Jahrhundert. In: Kämper / Warnke (Hg.): Diskurs interdisziplinär. Zugänge, Gegenstände, Perspektiven (= Diskursmuster – Discourse Patterns, hg. von Beatrix Busse/Ingo Warnke). Berlin: Akademie-Verlag.

Kollmeier, K. / Hoffmann, S.-L. (Hg.) (2012): Roundtable Discussion: Geschichtliche Grundbegriffe Reloaded? Writing the Conceptual History of the Twentieth Century. In: Contributions to the History of Concepts 7 (2), 78–128.

Kollmeier, K. / Hoffmann, S.-L. (Hg.) (2010): Zeitgeschichte der Begriffe? Perspektiven einer Historischen Semantik des 20. Jahrhunderts. Debatte, in: Zeithistorische Forschungen / Studies in Contemporary History 7 (1), 75–114. Online: http://www.zeithistorische-forschungen.de/16126041-Kollmeier-Hoffmann-1-2010.

Manning, C. D. / Raghavan, P. / Schütze, H. (2008): Introduction to Information Retrieval. Cambridge: Cambridge University Press.

McEnery, T. / Xiao, R. / Tono, Y. (2006): Corpus-Based Language Studies. An Advanced Resource Book (Routledge Applied Linguistics). London, New York: Routledge.

Navigli, R. (2009): Word sense disambiguation: A survey. ACM Computing Surveys (CSUR), 41 (2), 1–69.

Lloyd, S. P. (1957/1982): Least squares quantization in PCM. In: IEEE Transactions on Information Theory, 28, 129–137.

Lüdeling, A. / Kytö, M. (2008/9) (Hg.): Corpus Linguistics. An International Handbook. 2 Bände. Berlin, New York: de Gruyter.

Rohrdantz, C. et al. (2011): Towards Tracking Semantic Change by Visual Analytics. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, 305–310.

Storrer, A. (2011): Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In: Knapp et al. (Hg.): Angewandte Linguistik. Ein Lehrbuch. Tübingen: Francke, 216–239.