Big Data und Data Mining in den Digital Humanities

Big Data in den Geisteswissenschaften ermöglicht die Validierung und Extrahierung von Hypothesen aus großen Datenmengen wie sie es nie zuvor möglich war. In Sprachwissenschaften zum Beispiel können so Aussagen überprüft werden die sich nicht nur auf einzelne Textbeispiele beziehen sondern auf eine Verteilung über große Textmengen. So können Aussagen über Genres oder zeitspezifische Phänomene überprüft werden. In diesen Beitrag beschreiben wir Ergebnisse aus einem Projekt aus den Bereichen Korpuslinguistik und Data Mining. Auf Basis großer Datenmengen unterstützen wir mittels Data Mining Methoden linguistische Analysen.

In der Analyse von Sprache und allgemein in der Sprachforschung spielen große Textkorpora immer größere Bedeutung. Bei einem Korpus handelt es sich um eine Kollektion von Texten mit zusätzlichen Annotationen. Als Annotation verstehen wir eine Klassifizierung oder Beschreibung von Textelementen. Diese Annotationen können auf Textebene, Satzebene oder Wortebene vorliegen. Annotationen für Wörter können deren syntaktische Klasse sein oder verschiedene andere Schreibweisen des Wortes sein. Annotationen für Sätze können deren syntaktische Struktur als Parse-Baum darstellen. Annotation von ganzen Texten können Metainformationen wie deren Autoren, die Quelle, die Textsorte oder das Datum der Veröffentlichung sein.

Im Vergleich zu Anfragen an einen Korpus stellen Anfragen an eine Suchmaschine im Internet wie Google ungenügend Informationen zur Verfügung. So ist die Quellenlage bei einem Korpus geklärt. Dies ist bei der Sprachforschung besonders wichtig, da die Glaubwürdigkeit der Ergebnisse von den Quellen anhängt. Ferner können spezielle Features wie die oben genannten Annotationen berücksichtigt werden. Zum Beispiel kann man häufig auf linguistischen Korpora spezielle linguistische Features bei Anfragen berücksichtigen. So kann mit der Anfrage: http://www.dwds.de/?qu=bringen+with+%24p%3DVVFIN nach vorkommen des Verbs „bringen“ als finites Verb in den Texte gefragt werden.

Eine wichtige Aufgabe in den Sprachwissenschaften ist die Lesartendisambiguierung zur Erfassung von Mehrdeutigkeiten in der deutschen Sprache. Dabei werden für ein gegebenes Wort unterschiedliche Bedeutungen auf Basis des Kontextes in dem dieses Wort vorkommt bestimmt. Nachdem mögliche Bedeutungen ermittelt wurden, kann einem Vorkommen dieses Wortes dann eine Bedeutung, oder Lesart, zugewiesen werden. Ein Beispiel für solch eine Disambiguierung ist der Webservice Babelnet. Das Wort „Ampel“ kann man hier auf seinen unterschiedlichen Bedeutung hin untersuchen, wie man hier sehen kann: http://babelnet.org/exploreResult?word=Ampel&lang=DE

Diese Bestimmung der unterschiedlichen Bedeutungen bassiert auf den Vorkommen und dem Kontext des Wortes „Ampel“ in Wikipediaartikeln zum Beispiel. Dieser Datenbestand ist jedoch für sprachwissenschaftliche Untersuchungen zu ungenügend. So wird die Bedeutung des Wortes „Ampel“ als Blume nicht gefunden. Dies liegt an der Tatsache, dass „Ampel“ in dieser Bedeutung eher früher verwendet wurde als heute. In Wikipediaartikeln findet man

„Ampel“ hingegen nur in den heute gängigsten Bedeutungen als Lichtsignalanlage und als Lebensmittelampel.

Weiterhin sind die Entwicklung der Bedeutungen und die Verteilung dieser auf unterschiedlichen Genres hier nicht enthalten. Große Textkorpora wie sie von der Berlin Brandenburger Akademie der Wissenschaften (www.dwds.de) angeboten werden bieten hingegeben Texte aus unterschiedlichen Genres und Zeiten an. So kann man die Verteilung einer bestimmten Bedeutung eines Wortes in Zeitungsartikeln im Vergleich zur Belletristik untersuchen. Auch die Entwicklung der Bedeutungen über die Zeit ist so möglich.

In dem vom BmBF (Bundesministerium für Bildung und Forschung) geförderten Projekt KobRA http://www.kobra.tu-dortmund.de/mediawiki/index.php?title=Hauptseite werden unterschiedlichen Methoden zur Ermittlung und Zuweisung der verschiedenen Bedeutungen bestimmter Wörter mit Hilfe von Data Mining und Maschinellen Lernen entwickelt und evaluiert. Hierbei werden große Textkorpora wie das Deutsche Textarchiv (www.deutschestextarchiv.de) zusammen mit Informationen über die Textsorten und Veröffentlichungzeitpunkt mit berücksichtigt.

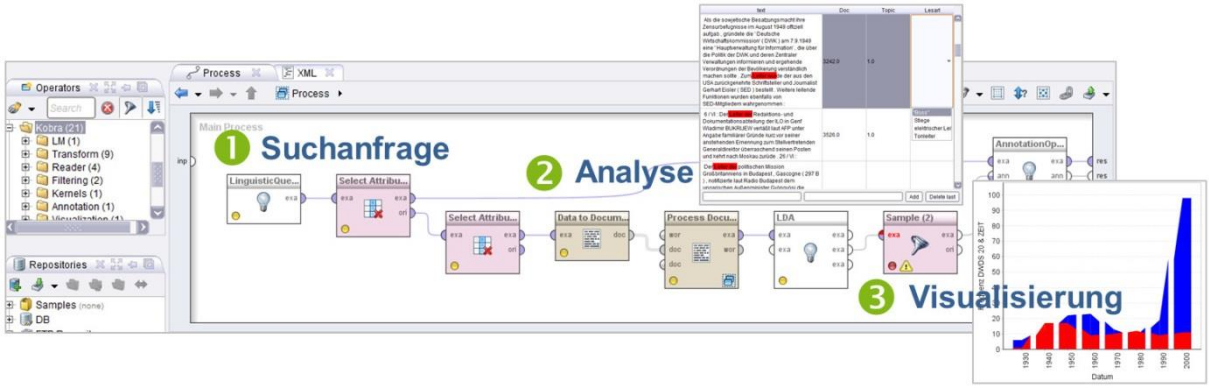

Abbildung 1: Visualisierung des Prozesses der Lesartendisambiguierung

Zur Bestimmung der unterschiedlichen Lesarten verwenden wir Topic Modelle wie Latent Dirichlet Allocation (Blei et al., 2002). Dabei wird ein probabilistisches Model ermittelt welches die Wörter auf eine vorgegebene Anzahl an möglichen Topics oder Lesarten verteilt. Ein im Rahmen des Projektes entwickeltes Tool ermöglicht es den Prozess der Disambiguierung eines Wortes komplett durchzuführen. Dieses Tool ist ein Plugin in dem bereits erfolgreich bestehenden Data Mining Werkzeug Rapidminer (https://rapidminer.com/). In Abbildung 1 ist ein Prozess zur Disambiguierung eines Wortes dargestellt. Zuerst werden Texte, die ein bestimmtes Wort enthalten (zum Beispiel Ampel), von einem Textkorpus extrahiert. Hierbei können auch die oben erwähnt linguistische Features mit angefragt werden. Die erhaltenden Textbeispiele aus dem Korpus können nun visuell untersucht und eventuell schon per Hand annotiert werden, so dass sie einer bestimmten Lesart angehören. Dies ist wichtig, da man so einen Goldstandard erhält an dem man die später automatisch ermittelten Lesarten bewerten kann. Nach einer Umwandlung der Texte in eine interne Repräsentation wird ein Topic Model berechnet. Dieses Model ermittelt die Wahrscheinlichkeiten, dass ein

bestimmtes Wort oder der ganzer Text zu einem bestimmten Topic (Lesart) gehört. Ein wichtiges Merkmal unserer Methoden ist, dass man die Möglichkeit hat die Entwicklung dieser gefundenen Lesarten über die Zeit und über Textsorten zu ermitteln und zu visualisieren.

Neben der visuellen Darstellung der Ergebnisse ermitteln wir ferner auch Gütemaße die die automatisch extrahierten Lesarten mit dem per Hand vorgegebenen Goldstandard vergleichen. Wir benutzen Normalized Mutual Information (NMI) und den F1 Score um die Lesarten aus dem Topic Model zu bewerten. NMI misst wie viele Texte mit gleicher Lesart im Goldstandard auch von unseren Topic Model der gleichen Lesart zugeordnet werden (Manning et al. 2008, p. 357f). Der F1 Score ist der relative Mittelwert aus den richtig der gleichen Lesart zugewiesenen Texte und der Anzahl aller dieser Lesart zugewiesenen Texte (Navigli et al. 2010).

Mit unseren Software Tool haben wir Experimenten zur Disambiguierung des Wortes „Leiter“ durchgeführt. Wir haben aus dem DWDS Kernkorpus (www.dwds.de) Sätze extrahiert, die das Wort „Leiter“ enthalten. Davon wurden 30 Prozent per Hand annotiert und zur Evaluierung verwendet. In der Tabelle in Abbildung 2 haben wir für die extrahierten Lesarten die Wörter aufgelistet, die die höchste Wahrscheinlichkeit haben in dieser Bedeutung zusammen mit „Leiter“ aufzutreten. Fernen haben wir in der Tabelle in Abbildung 3 die Wahrscheinlichkeiten aufgelistet, dass in einem bestimmten Genre eine dieser Lesarten verwendet wird.

Es wird jedem Text eine Bedeutung zugeordnet, die am wahrscheinlichsten ist, gegeben die Wörter in diesem Text. Mit dieser Zuordnung berechneten wir die NMI und den F1 Score. Für das ermittelte Topic Model ergibt sich eine NMI von 0,2573 und ein F1 Score von 0,7416. Wenn wir die Tabelle in Abbildung 2 anschauen, sehen wir dass die Bedeutung 2 wahrscheinlich „Leiter“ in der Bedeutung von Trittleiter meint. Die Bedeutungen 3 und 4 hingegen beinhalten eher Begriffe die auf den „Leiter“ als politischen Leiter deuten. Wenn wir nun in der Tabelle in Abbildung 3 die Verteilung der Bedeutungen über die Genres anschauen, sehen wir dass „Leiter“ in der Bedeutung als Trittleiter eher in Belletristik auftaucht als zum Beispiel in Zeitungsartikeln.

Bedeutung 1	**Bedeutung 2**	**Bedeutung 3**	**Bedeutung 4**
Musik	Stehen	DDR	Regierung
Berlin	Sehen	SED	Haben
Professor	Oben	Partei	Berlin
Komposition	Oberhalb	Politisch	ZK

Abbildung 2: Häufigste Wörter der ermittelten Bedeutungen

Leiter	**Bedeutung 1**	**Bedeutung 2**	**Bedeutung 3**	**Bedeutung 4**
Belletristik	0,0117707368	0,5777281878	0,0781636158	0,0706102457
Gebrauchsliteratur	0,059760642	0,1284352487	0,6671553638	0,0917891152

Wissenschaft	0,8194956525	0,0792926965	0,012734051	0,010994093
Zeitungen	0,1089729687	0,2145438669	0,2419469694	0,826606546

Abbildung 3: Wahrscheinlichkeit, dass ein Wort in mit ein bestimmten Bedeutung in einem Genre vorkommt

Mit den von uns entwickelten Methoden wird es künftig möglich sein quantitative empirische Untersuchungen auf großen Textkorpora durchzuführen, zu visualisieren und zu evaluieren. Die Integration von speziellen linguistischen Features und Metainformationen über die Texte ermöglichen es komplexe linguistische Analysen durchzuführen. In der Zukunft wollen wir eine noch nahtlosere Anbindung an externe Informationsquellen wie Wortnetzen (www.wordnet.de) und Baumdatenbanken wie die Tüba-DZ (http://www.sfs.uni-tuebingen.de/de/ascl/ressourcen/corpora/tueba-dz.html) in die Lesartendisambiguierung ermöglichen. Dadurch wollen wir nicht nur große Datenmengen nutzen sondern auch heterogene Datenquellen einbauen.

Blei, David M., Ng, Andrew Y., and Michael I. Jordan (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993-1022.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (EMNLP '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 116-126.