Automatische Textanalysen in der Geschichtswissenschaft – Auswertung, Interpretation und Relevanz

Fiedler, Maik
fiedler@gei.de
Georg Eckert Institut für internationale Schulbuchforschung, Deutschland

Weiß, Andreas
weiss@gei.de
Georg Eckert Institut für internationale Schulbuchforschung, Deutschland

Heuwing, Ben
heuwing@uni-hildesheim.de
Institut für Informationswissenschaft & Sprachtechnologie, Universität Hildesheim

Schnober, Carsten
schnober@ukp.informatik.tu-darmstadt.de
Ubiquitous Knowledge Processing Lab, Deutsches Institut für Internationale Pädagogische Forschung / Technische Universität Darmstadt

Motivation und Fragestellung

In jedem Digital-Humanities-Projekt (DH) stellt sich die Frage von neuem: Welche „Relevanz haben die Modellierung, Vernetzung und Visualisierung für die Geistesartefakte selbst und für den Gewinn reproduzierbarer wissenschaftlicher Erkenntnisse über sie? Im Projekt „Welt der Kinder“ (WdK) wurden diese Punkte mit Hilfe von Topic Modeling und Text-Mining-Werkzeugen mit in der Geschichtswissenschaft anerkannten Thesen in einem kontrollierten Verfahren überprüft. Es handelt sich bei WdK mit seinem repräsentativen Textkorpus von über 3000 historischen Schulbüchern um ein bisher weltweit einzigartiges Projekt, das für künftige ähnliche Vorhaben vorbildhaft sein will.

Aus Sicht der klassischen Geschichtswissenschaften gibt es bei der Bearbeitung großer Datenmengen häufig Argwohn gegenüber der Sekundäranalyse maschinell generierter Ergebnisse, verstärkt durch mangelndes Wissen über fachfremde Methodik. Dies lässt die Ergebnisse der DH oft als zweifelhaft oder nicht neuwertig erscheinen. Zusätzlich können Verzerrungen durch die Zusammensetzung einer Textsammlung entstehen, durch die Dokumentenauswahl und des zu analysierenden Vokabulars sowie aus den darauf aufbauenden Aggregierungen und Visualisierungen (Chuang et al. 2012). Große Datenmengen erfordern ein anderes Vorgehen bei der Auswertung als die traditionell in den Geschichtswissenschaften üblichen Verfahren. Die Methode der automatischen Textanalyse stellen trotzdem eine durch die Forschungsziele beeinflusste subjektive Sichtweise auf die vorhandenen Daten dar (DiMaggio et al. 2013). Wir zeigen an Hand eines in WdK vorgenommen Validierungsexperiments, welche Aushandlungsprozesse notwendig waren, um nachnutzbare und nachvollziehbare Ergebnisse zu erhalten.

Für die Meta-Analyse von klassischen und digitalen geschichtswissenschaftlichen Herangehensweisen ist die Beantwortung folgender Fragen prioritär:

Erstens) Wie können auf klassischem Weg erbrachte Ergebnisse für die DH so codifiziert werden, dass sie nicht nur für Menschen interpretierbar, sondern auch durch die digitalen Werkzeuge reproduzierbar sind? Sinn dieses Verfahrens ist es, Versuchsanordnungen und Analysen so aufzubauen, dass diese nicht immer „bei Null“ beginnen müssen, sondern, wie ein klassischer Fachtext, anerkannte Annahmen und Erkenntnisse implizit transportieren und wiederholen.

Zweitens) Wie kann die Belastbarkeit von Ergebnissen, die mit Hilfe von Methoden der automatischen Textmodellierung auf einem umfangreichen Korpus erbracht worden sind, validiert werden?

Drittens) Wie kann man die Leistung digitaler Methoden für explorative Analysen anwenden, ohne auf ein bereits feststehendes Ziel hinzuarbeiten?

Viertens) Wie müssen die Versuchsanordnung und das Projekt aufgebaut werden, um den Daten zu vertrauen und sie interpretieren sowie kontextualisieren zu können?

Der Vortrag wird den Arbeitsprozess (interdisziplinäre Arbeit an historischen Thesen mit Hilfe digitaler Tools) analysieren, die verschiedenen fachspezifischen Methoden problematisieren sowie schlaglichtartig Wege beleuchten, die zu möglichen Antworten auf die gestellten Fragen führen können.

Werkzeuge

Die Grundlage der Topic-Modelling-basierten Analyse besteht auf im Bereich DH etablierter Methoden wie LDA (*Latent Dirichlet Allocation*; Blei et al. 2003). Dieses Verfahren ordnet Begriffe auf Basis von Kookkurrenz und statistischen Analysen einander zu und extrahiert Topics in Form gewichteter Wortlisten. Diese ergeben für menschliche Benutzer interpretierbare Listen, und erlauben eine automatische Inferenz von Topic-Verteilungen innerhalb eines Dokuments.

Die Validierungsstudie wurde mit einem interaktiven Prototyp durchgeführt, der die Texte im Korpus und Statistiken über die Ergebnismengen zugänglich

macht. Suchanfragen können sich auf Metadaten – beispielsweise Jahr und Ort der Veröffentlichung oder Schultyp – Termanfragen und Topic-Verteilungen beziehen. Ergebnisse werden mit Statistiken zur Topic-Intensität und relativen Dokumentenhäufigkeit im Zeitverlauf ausgegeben.

Vorgehen bei der Validierung:

Belastbarkeitsüberprüfungen bauen Vertrauen in datenbasierte, historische Schlussfolgerungen und Annahmen auf. So wird überprüft, ob die statistischen Modelle existierende Erkenntnisse mehrheitlich bestätigen, und als wie zuverlässig bestätigende oder widerlegende Ergebnisse eingeschätzt werden (DiMaggio et al. 2013; Evans 2014). Die im Experiment bearbeiteten historischen Thesen stellten Sachverhalte dar, die sich quantitativ überprüfen lassen, etwa durch den Vergleich von Topic-Verteilungen (Newman / Block 2006; Yang et al. 2011), und im Nachhinein von Experten für das jeweilige Fachgebiet in Hinblick auf ihre Plausibilität überprüft werden.

Für die Validierungsstudie wurden zu überprüfende Thesen vorab definiert, um Abweichungen von der ursprünglichen Fragestellung zu dokumentieren. Sie sind repräsentativ für reale historische Fragestellungen im Rahmen des Projektes (Kolonien und Auswanderung; Französische Revolution und Befreiungskriege; deutsche Kriegsflotte). Dabei wurden in einem ersten Schritt Begrifflichkeiten und Interpretationen der Fragestellungen in interdisziplinären Arbeitsgruppen diskutiert, um fachliche Verständnisschwierigkeiten auszuräumen. Da die Thesen erschöpfend und präzise mit den vorhandenen Werkzeugen untersucht wurden, bilden auch die Auswertungsstrategien mögliche Vorgehensweisen für die Überprüfung bereits vorliegender Hypothesen ab.

Auswertung

Bei der Analyse der Thesen zeigten sich unterschiedliche Strategien für die einzelnen Schritte der Auswertung. Wichtig hierbei war, ob unterschiedliche Herangehensweisen, vergleichbare Ergebnisse reproduzierten. Die Ergebnisse der einzelnen Arbeitsgruppen widersprachen einander an wenigen Stellen, und gegebenenfalls primär in ihrer Bewertung der Verlässlichkeit der Ergebnisse. Die vorgegebenen geschichtswissenschaftlichen Thesen wurden in den Versuchen mit Topic-Modellen größtenteils bestätigt und zusätzlich mittels Termanfragen validiert.

Das Vorgehen bei den Topic-Modelling-basierten Analysen beinhaltete im ersten Schritt eine Suche nach relevanten Topics an Hand einzelner Terme. Dabei zeigte sich, dass die Topics in Modellen mit einer manuell überschaubaren Topic-Anzahl (50, 100, 200) für spezielle historische Forschungsfragen zu allgemein oder auch zu spezifisch ausfielen. Teilweise wurden daraufhin die Thesen stellvertretend an Hand thematischer Teilgebiete oder übergeordneter Themen untersucht.

Für eine höhere Genauigkeit wurden auch Kombinationen aus Termsuche und Dokumentenfiltern auf Basis automatisch generierter Topics eingesetzt. Für eine Bewertung der Abfragegenauigkeit wurden manuelle Inspektionen der relevantesten Trefferdokumente durchgeführt und Anfragen iterativ neu formuliert. Um für die Validierung eine Vergleichsebene bereitzustellen, wurden zusätzliche Analysen nur auf der Grundlage manuell und mittels historischen Vorwissens gewählter Terme durchgeführt.

Schlussfolgerungen

Zusammengefasst kann zwischen zwei grundlegenden Vorgehensweisen unterschieden werden. In der ersten Variante werden die aufgestellten Thesen konfirmatorisch überprüft. Diese werden dafür formalisiert und in Form von Suchanfragen und zu erwartenden Ergebnissen operationalisiert. Die Ergebnisse werden dann vor allem hinsichtlich der erwarteten Zeitverläufe und relativen Unterschiede zwischen Untermengen interpretiert.

Die explorative Herangehensweise an die Datenanalyse berücksichtigt dagegen auch andere Hinweise aus den Ergebnissen, und sucht nach Erklärungen für beobachtete Auffälligkeiten. Die Aussagekraft der Ergebnisse kann dabei jedoch dadurch eingeschränkt werden, dass die untersuchten Thesen erst mit Kenntnis der Daten formuliert worden sind. Eine Strategie, um diese Unsicherheit auszugleichen, besteht darin, Evidenz für eine Aussage mit mehreren unterschiedlichen Vorgehensweisen zu sammeln.

Diese Ergebnisse zeigen den potentiellen Mehrwert von DH an, da mit Hilfe computerlinguistischer und informationswissenschaftlicher Methoden klassische Thesen aus der Geschichtswissenschaft präzisiert werden konnten. Die Interpretation quantitativer Ergebnisse, etwa als Diagramm visualisiert, konnte sich nach Bedarf auf die vorab definierten Vorannahmen beschränken. Die Einbeziehung größerer zeitlicher Kontexte erforderte teilweise, die dargestellten Verläufe und Tendenzen mit verschiedenen Zeitspannen neu zu interpretieren. Als wichtige Vorgehensweise hat sich hier die Bildung eines gleitenden Durchschnitts über längere Zeiträume erwiesen, um thematische Tendenzen zuverlässiger interpretieren zu können.

Als wichtiger Faktor stellte sich auch die Qualität der OCR-Digitalisierung heraus. Bei Daten aus historischen Quellen (Schriftbild Sütterlin / Fraktur) werden auch mit aktueller Technologie aufgrund der verwendeten Schriftarten teilweise über 10 Prozent der Zeichen falsch erkannt, was bei der Auswertung der maschinell generierten Topics durch die Benutzer zu Problemen bei der Interpretation und Weiterverwendung führt. Daher muss die Frage gestellt werden, wie Daten zukünftig in den Vorverarbeitungsschritten aufbereitet werden, damit

Topic Modelling und andere automatische Methoden zu hilfreichen und interpretierbaren Ergebnissen führen.

Neben der Einbeziehung von Topic Models, die auf unterschiedliche Perspektiven optimiert wurden, werden im Rahmen des Projektes andere Herangehensweisen an die statistische Textmodellierung, wie z. B. Clustering-Verfahren, in Hinblick auf ihre Anwendbarkeit und Robustheit vergleichend evaluiert. In diesem Zusammenhang ist es wichtig, thematisch relevante Topics einfach auffindbar zu machen und sie für Anfragen kombinieren zu können. Des Weiteren sollten Topics geordnet nach Themen oder Diskursfeldern und / oder -strängen präsentiert werden sowie in einer leicht lesbaren Anzeige deren synchrone und diachrone Verteilungen herausstellen, wobei Ungleichverteilungen innerhalb der Untersuchungsmenge und die Zuverlässigkeit statistischer Aggregierungen deutlich gemacht werden müssen.

Bibliographie

Blei, David. M. / Ng, Andrew. Y. / Jordan, Michael I. (2003): "Latent Dirichlet allocation", in: *Journal of Machine Learning Research* 3: 993–1022.

Chuang, Jason / Ramage, Daniel / Manning, Chistopher / Heer, Jeffrey (2012): "Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis", in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *CHI '12*. New York, NY, USA: ACM 443–452.

DiMaggio, Paul / Nag, Manish / Blei, David (2013): "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding", in: *Poetics* 41, 6: 570–606.

Evans, Michael S. (2014): "A Computational Approach to Qualitative Analysis in Large Textual Datasets", in: *PLoS ONE*9, 2, e87908.

Kaplan, Frédéric (2015): "A map for Big Data research in Digital Humanities", in: *Frontiers in Digital Humanities* 2, 1: http://journal.frontiersin.org/article/10.3389/fdigh.2015.00001/abstract [letzter Zugriff 08. Januar 2016].

Newman, David J. / Block, Sharon (2006): "Probabilistic topic decomposition of an eighteenth-century American newspaper", in: *Journal of the American Society for Information Science and Technology* 57, 6: 753–767.

Yang, Tze-I. / Torget, Andrew J. / Mihalcea, Rada (2011): "Topic modeling on historical newspapers", in: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics 96–104.