Entitäten als Topic Labels: Verbesserung der Interpretierbarkeit und Evaluierbarkeit von Themen durch Kombinieren von Entity Linking und Topic Modeling

Lauscher, Anne
anne@informatik.uni-mannheim.de
Universität Mannheim, Deutschland

Nanni, Federico
federico@informatik.uni-mannheim.de
Universität Mannheim, Deutschland

Ponzetto, Simone Paolo
simone@informatik.uni-mannheim.de
Universität Mannheim, Deutschland

Im letzten Jahrzehnt haben Wissenschaftler aus dem Bereich der Geisteswissenschaften zunehmend mit verschiedenen Text Mining-Techniken zur Exploration großer Textkorpora experimentiert. Angefangen bei Kookkurrenz-basierten Verfahren (Buzydlowski, White und Lin 2002) über automatische Keyphrase Extraktion (Hasan, Saidul und Ng 2014) ziehen sich die angewandten Techniken bis hin zu Sequence Labeling Algorithmen, wie zum Beispiel im Falle von Named-Entity Recognition (Nadeau und Sekine 2007). Aus diesen vielfältigen Techniken bedienten sich die Forscher in den letzten Jahren vor allem des Latenten Dirichlet Allokation (LDA) Topic Model Algorithmus (Blei, Ng und Jordan 2003) (Meeks und Weingart 2012). Oftmals betonten Wissenschaftler dessen Potential für Serendipität (Alexander et al. 2014) und für Analysen im Bereich des Distant Reading (Leonard 2014; Graham, Milligan und Weingart 2016), also Studien, die über reine Textexploration hinausgehen.

In den letzten Jahren wurde LDA in den Digitalen Geisteswissenschaften intensiv angewandt, obwohl bekannt ist, dass die damit erzielten Ergebnisse schwierig zu interpretieren (Chang et al. 2009; Newman et al. 2010) und dass die Möglichkeiten, deren Qualität zu evaluieren, stark begrenzt sind (Wallach et al. 2009). Die direkte Konsequenz daraus ist, dass Wissenschaftler im Bereich der Geisteswissenschaften momentan in einer Situation feststecken, in der sie Topic Models weiterhin anwenden, da sie Methoden dieser Art benötigen, aber auch gleichzeitig nur wenig neues geisteswissenschaftliches Wissen ableiten können, weil die erzielten Ergebnisse bereits intrinsisch begrenzt sind (Nanni, Kümper und Ponzetto 2016). Diese Situation ist vor allem darauf zurückzuführen, dass große Korpora bestehend aus Primärquellen nun zum ersten Mal digital verfügbar sind.

Von dieser Grundsituation ausgehend wollen wir dieses komplexe Problem bewältigen, indem wir zwei spezifische und integrierte Lösungen zur Verfügung stellen. Als erstes bieten wir eine neue Methode zur Exploration von Textkorpora, die Topics erzeugt, welche leichter zu interpretieren sind als traditionelle LDA Topics. Dies erreichen wir durch die Kombination zweier Techniken, nämlich Entity Linking und Labeled LDA. Unsere Methode identifiziert in einer Ontologie eine Serie beschreibender Labels für jedes Dokument in einem Korpus. Daraufhin wird für jedes der identifizierten Labels ein Topic erzeugt. Durch die daraus resultierende direkte Beziehung zwischen Topic und Label wird die Interpretation des Topics stark vereinfacht und durch die Ontologie im Hintergrund wird die Ambiguität der Labels vermindert. Da unsere Topics mit einer limitierten Anzahl an klar umrissenen Labels beschrieben werden, fördern sie die Interpretierbarkeit und die Anwendung der Ergebnisse als quantitativ grundierte Argumente in der geisteswissenschaftlichen Forschung.

Da es äußerst wichtig ist, die Qualität der Ergebnisse zu bestimmen, stellen wir zweitens eine dreischrittige Evaluationsplattform zur Verfügung, die die Ergebnisse unseres Ansatzes als Input verwendet und eine umfangreiche quantitative Analyse ermöglicht. Dies gestattet den nutzenden Wissenschaftlern aus den Digitalen Geisteswissenschaften, einen Überblick über die Ergebnisse der einzelnen Schritte der Pipeline zu erhalten und stellt Forschern im Natural Language Processing (NLP) eine Serie von Baselines zur Verfügung, die sie zur Verbesserung jedes Schrittes der vorgestellten Methodik benutzen können.

Wir illustrieren das Potenzial dieses Ansatzes durch dessen Anwendung zur Bestimmung der relevantesten Topics in drei verschiedenen Datensätzen. Der erste Datensatz besteht aus der gesamten Transkription der Reden aus dem fünften Mandat des Europäischen Parlaments (1999-2004). Dieses Korpus (van Aggelen et al. 2016) wurde für Forschung im Bereich der Computational Political Science bereits intensiv eingesetzt (Hoyland und Godbout 2008; Proksch und Slapin 2010; Høyland et al. 2014) und hat enormes Potential für zukünftige politikgeschichtliche Forschungen. Das zweite Korpus ist der sogenannte Enron-Datensatz. Es handelt sich dabei um eine große Datenbank mit über 600.000 E-Mails, die von 158 Mitarbeiten der Enron Corporation erstellt und die später durch die Federal Energy Regulatory Commission während der Untersuchungen nach dem Zusammenbruch des Unternehmens akquiriert wurden. In den letzten zehn Jahren hat die NLP-Community diesen Datensatz unter Anwendung von netzwerk- und inhaltsbasierten Analysen intensiv untersucht. Unser Ziel ist es hierbei, die Qualität unseres Ansatzes anhand eines hochtechnischen und komplexen Datensatzes einer spezifischen Art (E-Mail),

die in zukünftigen historischen Untersuchungen immer wichtiger werden wird, zu beleuchten. In Verbindung damit wurde als drittes Korpus der Hillary Clinton E-Mail-Datensatz ausgewählt. Er repräsentiert eine Kombination der beiden vorherigen Datensätze, da es sich um kurze Korrespondenzen via E-Mail handelt, die sich jedoch mehrheitlich auf politische Themen fokussieren.

Vor über einem Jahrzehnt hat Dan Cohen (2006) bereits vorhergesehen, dass künftige Politikhistoriker in Anbetracht der Fülle an Quellen, die die öffentliche Verwaltung uns in den kommenden Jahrzehnten hinterlassen wird, auf ein Problem stoßen werden. Unsere Studie möchte ein allererster experimenteller Ansatz zu sein, diese neuen Korpora von Primärquellen zu bewältigen und Historiker im digitalen Zeitalter mit einer feinkörnigeren Lösung zur Textexploration als mittels traditionellen LDAs auszustatten.

Bibliographie

Alexander, Eric / Kohlmann, Joe / Valenza, Robin / Witmore, Michael / Gleicher, Michael (2014): „Serendip: Topic model-driven visual exploration of text corpora“, in: *IEEE VAST* 173–182.

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

Buzydlowski, Jan W. / White, Howard D / Lin, Xia (2002): „Term co-occurrence analysis as an interface for digital libraries“, in: *Visual interfaces to digital libraries*. Springer 133–144.

Chang, Jonathan / Gerrish, Sean / Wang, Chong / Boyd-Graber, Jordan L. / Blei, David M. (2009): „Reading tea leaves: How humans interpret topic models“, in: *NIPS* 288–296.

Cohen, Dan (2006): *When machines are the audience*.

Graham, Shawn / Milligan, Ian / Scott Weingart (2016): *Exploring big historical data: The historian's macroscope*. Imperial College Press.

Hasan, Kazi Saidul / Ng, Vincent (2014): „Automatic keyphrase extraction: A survey of the state of the art“, in: *Proceedings of ACL-2014* 1262–1273.

Høyland, Bjørn / Godbout, Jean-Franc#ois (2008): *Lost in translation? Predicting party group affiliation from European parliament debates*. Unveröff. Manuskript.

Høyland, Bjørn / Godbout, Jean-Franc#ois / Lapponi, Emanuele / Velldal, Erik (2014): „Predicting party affiliations from European parliament debates“, in: *ACL 2014 Workshop on Language Technologies and Computational Social Science* 56–60.

Leonard, Peter (2014): „Mining large datasets for the humanities“ in: *IFLA WLIC* 16–22.

Meeks, Elijah / Weingart, Scott B. (2012): „The digital humanities contribution to topic modeling“, in: *Journal of Digital Humanities* 2 (1): 1–6.

Nadeau, David / Sekine, Satoshi (2007): „A survey of named entity recognition and classification“, in *Lingvisticae Investigationes* 30 (1): 3–26.

Nanni, Federico / Kümper, Hiram / Ponzetto, Simone Paolo (2016): „Semi-supervised textual analysis and historical research helping each other: Some thoughts and observations“ in: *International Journal of Humanities and Arts Computing* 10 (1): 63–77.

Newman, David / Lau, Jey Han / Grieser, Karl / Baldwin, Timothy (2010): „Automatic evaluation of topic coherence“, in: *HLT-NAACL* 100–108.

Proksch, Sven-Oliver / Slapin, Jonathan B. (2010): „Position taking in European parliament speeches“, in: *British Journal of Political Science* 40 (3): 587–611.

van Aggelen, Astrid / Hollink, Laura / Kemman, Max / Kleppe, Martijn / Beunders, Henri (2016): „The debates of the European parliament as linked open data“, in: *Semantic Web* (Preprint) 1–10.

Wallach, Hanna M. / Murray, Iain / Salakhutdinov, Ruslan / Mimno, David (2009): „Evaluation methods for topic models“, in: *ICML* 1105–1112.