Erweiterbare, interaktive Softwareplattform für die Anwendung von Sprachtechnologie in großen Textkorpora zur Unterstützung von Such- und Analyseworkflows in den Digital Humanities

Fynn Petersen-Frey, Universität Hamburg, Deutschland

Pre-Print zur DHd2023

Motivation und Einleitung

In Wissenschaftsdisziplinen wie den digitalen Geistes- und Sozialwissenschaften „Digital Humanities“ (DH), besteht ein großes Interesse daran, umfangreiche Mengen von Text, z.B. aus historischen Zeitschriften (Purschwitz 2018, 109–142), sozialen Medien (Stier u. a. 2017, 1365–1388) oder Zeitungen, hinsichtlich verschiedener Fragestellungen auszuwerten. Dabei ist oft eine Kombination von qualitativen Analyseschritten mit quantitativen Auswertungen gefragt (Stulpe und Lemke 2016, 17–61).

Eine ausschließlich manuelle Bearbeitung ist angesichts des Umfangs und des daraus resultierenden Aufwands oftmals ausgeschlossen. In diesen Fällen ermöglicht es eine (teil-)automatisierte computerlinguistische Verarbeitung dennoch Analyse und Auswertung durchzuführen.

Ziel dieses Dissertationsvorhabens ist es daher geeignete Methoden zu entwickeln und in eine moderne Sprachtechnologieplattform zu integrieren, die es Wissenschaftlern aus den DH ermöglicht, selbstständig große Textkorpora mit Hilfe (teil-)automatischer Verarbeitungsschritte aus der CL vorzubereiten und hinsichtlich vielfältiger Fragestellungen auszuwerten. Dies umfasst lexikalische wie semantische Suchfunktionen, intuitive Annotationsfunktionen, AI- bzw. Human-in-the-Loop Funktionalitäten zur (teil-)automatischen Skalierung von Annotationen auf große Korpora, über klassische syntaktische Anreicherungen hinausgehende inhaltliche NLP-Methoden wie Koreferenzausflösung und Zitaterkennung sowie qualitative und quantitative Analysemöglichkeiten.

Forschungsstand

Verwandte Software aus der Computerlinguistik wie *brat* (Stenetorp u. a. 2012, 102–107), *WebAnno* (Eckart de Castilho u. a. 2016, 76–84), *INCEpTION* (Klie u. a. 2018, 5–9) oder *TextAnnotator* (Abrami u. a. 2020, 891–900) sind vorrangig für linguistische Annotationen gedacht, um annotierte Korpora zu erstellen; weniger jedoch für die inhaltliche Bearbeitung einer DH-Forschungsfrage mittels Suche, Annotation, Aggregation und Analyse.

Für spezifische Recherchezwecke existieren computerlinguistische Anwendungen wie *SiNLP* (Crossley u. a. 2014, 511–534) zur Diskursanalyse, *ALCIDE* (Moretti u. a. 2016, 100–112) zur Analyse von historischem und politischem Diskurs, *new/s/leak* (Wiedemann u. a. 2018, 313–322) zum Entdecken berichtenswerter Geschehnisse in großen Textkorpora sowie *LawStats* (Ruppert u.

a. 2018, 212–222) zur Suche und Analyse von Revisionen des Bundesgerichtshofs. Diese Anwendungen sind jedoch weniger geeignet andere Fragestellungen zu bearbeiten.

In den Sozialwissenschaften gängige qualitative Analysetools wie *MaxQDA* oder *atlas.ti* verfügen über Annotations- und qualitative Analysefunktionen, sind aber proprietär, erlauben kaum Kollaboration und bieten keine modernen NLP-Methoden.

Recogito (Simon u. a. 2017, 111–132) und *CATMA* (Gius u. a., 2022) sind intuitive Annotationsanwendungen für die DH mit Kollaborationsfunktionen. *Recogito* setzt einen Fokus auf Orte und Integration in eine Karte mittels automatischer Erkennung von Entitäten. *CATMA* bietet konfigurierbare Annotationsschemata und vielfältige Analysefunktionen. Beide verfügen jedoch nur über eine bescheidene bzw. keine Suche, wenig Unterstützung für große Korpora und keine Möglichkeit zur Skalierung manueller Annotationen.

WebLicht (Hinrichs u. a. 2010, 25–29) integriert enorm viele NLP-Module, die zu individuellen Pipelines zusammengefügt werden können. *Nopaque* (Universität Bielefeld 2022) unterstützt historische Dokumente mittels OCR und syntaktische Analysen anhand Keyword-In-Context-Suche auf Basis gängiger NLP-Modelle und individueller NLP-Pipelines. Beiden fehlen jedoch Annotationsmöglichkeiten, Kollaborationsfunktionen, Dokumentensuche sowie Möglichkeiten zur quantitativen Analyse großer Korpora.

ILCM (Niekler u. a. 2018, 1313–1319) ist eine Textmining-Umgebung für die datengetriebene Forschung auf der großen Textmengen mittels statistischer Analysen. Es fehlen jedoch moderne NLP-Modelle sowie interaktive Funktionalitäten wie Annotation oder Suche.

Forschungsvorhaben und Forschungsfragen

Das Forschungsvorhaben steht unter der übergreifenden Forschungsfrage: Wie kann eine digitale Arbeitsumgebung entwickelt werden, welche undogmatisch die Anwendung von DH-Methoden durch moderne NLP-Methoden für Suche, Annotation, Skalierung, Aggregation und Analyse auf großen Korpora unterstützt?

Das Ziel ist es geeignete Methoden für eine Sprachtechnologieplattform zu entwerfen, die intuitiv aus den DH genutzt werden kann, um manuelle Arbeit mit Textdokumenten automatisch auf große Korpora zu skalieren. Dies umfasst u. a. eine semantischen Suche um ähnliche Aussagen zu einer bestimmten Textpassage im gesamten Korpus zu finden, die Möglichkeit gefundenen Textpassagen qualitativ manuell zu analysieren oder automatisch zu aggregieren für quantitative Auswertungen. Im Zusammenspiel von Entity-Linking, Koreferenzauflösung, Zitaterkennung und Auffinden ähnlicher Textpassagen soll eine Skalierung manueller Annotation bzw. Kodierung von Textspannen auf den gesamten Korpus ermöglicht werden, indem Annotationen einzelner Textstellen auf passende Stellen gesamten Korpus angewandt wird.

Wie können dem aktuellen Stand der Forschung entsprechende computerlinguistische Methoden undogmatisch und intuitiv für die Bearbeitung verschiedener DH-Fragestellungen nutzbar gemacht werden?

Zum aktuellen Stand der Forschung zählen kontextualisierte Embeddings wie *BERT* (Devlin u. a. 2019, 4171–4186), die jedem Wort und (Ab-)Satz eine Bedeutung in Abhängigkeit des Kontexts anhand der Position in einem hochdimensionalen Vektorraum zuordnen. Wie lässt sich auf dieser Basis eine semantische Ähnlichkeitssuche zum Auffinden verwandter Aussagen mit variierender Länge entwickeln? Dabei ist zu klären, welche Lösungen für die rechenintensiven Operationen bei der Ähnlichkeitssuche mit großen Korpora skalieren.

Um diese NLP-Methoden nutzbar zu machen, werden sie in eine webbasierte Benutzeroberfläche integriert, die das Durchsuchen, Annotieren, Skalieren, Aggregieren und Auswerten großer Korpora mittels der zuvor beschriebenen computerlinguistischen Funktionalitäten unterstützt. Ferner wird geeignete Softwarearchitektur entworfen, welche die Integration bestehender und zukünftiger CL-Softwarebibliotheken ermöglicht.

Bibliographie

Abrami, Giuseppe, Manuel Stoeckel und Alexander Mehler. 2020. “TextAnnotator: A UIMA Based Tool for the Simultaneous and Collaborative Annotation of Texts”. English. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, S. 891–900. isbn: 979-10-95546-34-4. url: https://www.aclweb.org/anthology/2020.lrec-1.112.

Crossley, Scott A., Laura K. Allen, Kristopher Kyle und Danielle S. McNamara. 2014. “Analyzing Discourse Processing Using a Simple Natural Language Processing Tool”. In: *Discourse Processes 51.5-6*, S. 511–534. doi: 10.1080/0163853X.2014.910723.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, S. 4171–4186. doi: 10.18653/v1/N19- 1423. url: https://www.aclweb.org/anthology/N19-1423.

Eckart de Castilho, Richard, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank und Chris Biemann. 2016. “A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures”. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee, S. 76–84. url: https://www.aclweb.org/anthology/W16-4011.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, und Jan Horstmann. 2022. “CATMA”. Zenodo, url: https://doi.org/10.5281/zenodo.6419805.

Hinrichs, Erhard W., Marie Hinrichs and Thomas Zastrow. 2010. “WebLicht: Web-Based LRT Services for German”. In: Proceedings of the ACL 2010 System Demonstrations. S. 25–29. url: http://www.aclweb.org/anthology/P10-4005

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho und Iryna Gurevych. 2018. “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation”. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Association for Computational Linguistics*, S. 5–9. url: http://tubiblio.ulb.tu-darmstadt.de/106270/.

Moretti, Giovanni, Rachele Sprugnoli, Stefano Menini und Sara Tonelli. 2016. “ALCIDE: Extracting and visualising content from large document collections to support humanities studies”. In: *Knowledge-Based Systems 111*, S. 100–112. issn: 0950-7051. doi: https://doi.org/10.1016/j.knosys.2016.08.003. url: http://www.sciencedirect.com/science/article/pii/S0950705116302635.

Niekler, Andreas, Arnim Bleier, Christian Kahmann, Lisa Posch, Gregor Wiedemann, Kenan Erdogan, Gerhard Heyer und Markus Strohmaier. 2018. “ILCM - A Virtual Research Infrastructure for Large-Scale Qualitative Data”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), S. 1313–1319 url: https://www.aclweb.org/anthology/L18-1209.

Purschwitz, A. (2018). “Netzwerke des Wissens – Thematische und personelle Relationen innerhalb der halleschen Zeitungen und Zeitschriften der Aufklärungsepoche (1688-1818)”. In: *Journal of Historical Network Research* 2 (1), S. 109–142.

Ruppert, Eugen, Dirk Hartung, Phillip Sittig, Tjorben Gschwander, Lennart Rönneburg, Tobias Killing und Chris Biemann. 2018. LawStats – Large-Scale German Court Decision Evaluation Using Web Service Classifiers. In: *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018*, Hamburg, Germany, August 27–30, 2018, Proceedings. Springer-Verlag, Berlin, Heidelberg, 212–222. doi: 10.1007/978-3-319-99740-7_14

Simon, Rainer, Elton Barker, Leif Isaksen und Pau De Soto CaÑamares. 2017. “Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2”. In: Journal of Map & Geography Libraries, 13:1, S. 111–132. doi: 10.1080/15420353.2017.1307303

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou und Jun’ichi Tsujii. 2012. “brat: a Web-based Tool for NLP-Assisted Text Annotation”. In: *Proceedings*

of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, S. 102–107. url: https://www.aclweb.org/anthology/E12-2021.

Stier, Sebastian, Lisa Posch, Arnim Bleier und Markus Strohmaier. 2017. “When populists become popular: comparing Facebook use by the right-wing movement Pegida and German political parties”. In: *Information, Communication & Society* 20, S. 1365–1388. doi: 10.1080/1369118X.2017.1328519.

Stulpe, Alexander und Matthias Lemke. 2016. “Blended Reading”. In: *Text Mining in den Sozialwissenschaften: Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Hrsg. von Matthias Lemke und Gregor Wiedemann. Wiesbaden: Springer Fachmedien Wiesbaden, S. 17–61. isbn: 978-3-658-07224-7. doi: 10.1007/978-3-658-07224-7_2. url: https://doi.org/10.1007/978-3-658-07224-7_2.

Wiedemann, Gregor, Seid Muhie Yimam und Chris Biemann. 2018. “New/s/leak 2.0 – Multilingual Information Extraction and Visualization for Investigative Journalism”. In: *Social Informatics*. Hrsg. von Steffen Staab, Olessia Koltsova und Dmitry I. Ignatov. Cham: Springer International Publishing, S. 313–322. isbn: 978-3-030-01159-8.

Universität Bielefeld. 2022. “nopaque”. https://nopaque.uni-bielefeld.de (zugegriffen 12. Dezember 2022)