Pipelines für Natural Language Processing und digitale Literaturanalyse in spaCy

Varachkina, Hanna
hanna.varachkina@stud.uni-goettingen.de
Seminar für Deutsche Philologie, Georg-August-Universität Göttingen

Barth, Florian
florian.barth@uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Dönicke, Tillmann
tillmann.doenicke@uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Biermann, Johannes
johannes.biermann@gwdg.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Altmann, Friederike
friederike.altmann@stud.uni-goettingen.de
Seminar für Deutsche Philologie, Georg-August-Universität Göttingen

Neitzke, Thorben
thorben.neitzke@stud.uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Sporleder, Caroline
caroline.sporleder@cs.uni-goettingen.de
Göttingen Centre for Digital Humanities, Georg-August-Universität Göttingen

Kontext und Bedarf

In diesem halbtägigen Workshop stellen wir ein auf spaCy basierendes Pipeline-System für das Natural Language Processing (NLP) narrativer Texte vor und erproben mit den Teilnehmer*innen dessen praktische Anwendung, besonders im Hinblick auf Untersuchungsgegenstände der digitalen Literaturanalyse.

Die Analyse von literarischen Texten ist eine besondere Herausforderung für die automatische Sprachverarbeitung, da sie oft komplexe Interaktionen linguistischer Strukturen auf der syntaktischen, semantischen und pragmatischen Ebene betrifft. Für die Interpretation solcher Texte ist es zum Beispiel wichtig, neben traditionellen NLP-Verarbeitungsschritten wie Eigennamenerkennung, Sentiment-Analyse etc., auch komplexere Analysen durchzuführen, um z. B. die Sprechinstanzen im Text zu identifizieren, Bezüge zur realen Welt zu erkennen oder zeitliche Strukturen im Text zu analysieren. Auf der praktischen Ebene bedeutet dies, dass automatische Analysen in der digitalen Literaturwissenschaft in der Regel die (oft komplexe) Kombination mehrerer basaler Sprachverarbeitungswerkzeuge auf Token-, Teilsatz-, Satz- und Passagen-/Diskursebene erfordert. Dies ist in der Praxis nicht immer trivial, z. B. weil Ein- und Ausgabeformate verschiedener Werkzeuge nicht kompatibel sind.

Bibliotheken wie spaCy stellen sehr umfassende Sammlungen von Sprachverarbeitungswerkzeugen zur Verfügung, können potenzielle Nutzer*innen durch ihre Fülle und Heterogenität aber auch überfordern. Das vorgestellte Pipeline-System MONAPipe (Dönicke u. a. 2022) soll hier Abhilfe schaffen, indem es Werkzeuge für linguistische und literaturwissenschaftliche Analysen komfortabel bündelt und flexibel erweiterbar ist. Der Fokus liegt dabei auf narrativen Texten und auf typischen Anwendungsszenarien der digitalen Literaturanalyse.

Der Workshop vermittelt (i) die Grundlagen von spaCy und dessen Kernkomponenten (Tokenisierung, Lemmatisierung, Erkennung von Satz- und Teilsatzgrenzen, Dependency Parsing), (ii) demonstriert, wie MONAPipe an die eigenen Zwecke durch Custom-Komponenten angepasst werden kann, und versetzt (iii) die Teilnehmer*innen mit hands-on Praxisbeispielen in die Lage, die in MONAPipe integrierten Komponenten zur Erschließung der linguistischen und narrativen Struktur eines Textes im Rahmen eigener Projekte kompetent auszuwählen, anzuwenden, zu erweitern und die Ergebnisse zu beurteilen. Unter anderem behandeln wir die Erkennung von Named Entities (sowie das Linking zu Normdaten; vgl. Barth u. a. 2022), von Zeitformen (Dönicke 2020), Eventtypen (Vauth u. a. 2021) und Redeformen (direkte, indirekte, erlebte Rede; vgl. Brunner u. a. 2020), Animatheit (Tuggener u. Klenner 2014) sowie Sentiment- (Remus u. a. 2010) und Emotionsanalyse (Mohammad u. Turney 2013). Schließlich erproben wir mit den Teilnehmer*innen, wie die Wechselwirkung einzelner Komponenten von MONAPipe Muster in Erzähltexten aufdecken kann, die zur Modellierung komplexer linguistischer und narrativer Phänomene geeignet sind (z. B. Generalisierungen (Gödeke u. a. 2022) oder narrative Kommentare (Weimer u. a. 2022)).

Technische Voraussetzungen

Wir stellen Jupyter-Notebooks bereit, in denen MONAPipe und alle benötigten Dependencies vorinstalliert sind. Die Teilnehmer*innen benötigen Kenntnisse in Python; Erfahrung im Umgang mit Jupyter-Notebooks und der Unix-Kommandozeile ist hilfreich.

Zielpublikum

Der Workshop ist als Tutorial geplant und richtet sich an Literaturwissenschaftler*innen, Linguist*innen, DH-Forschende, und andere Personen, die an Textanalyse interessiert sind. Die Teilnehmer*innen bekommen die Möglichkeit, die Funktionalitäten von MONAPipe auszuprobieren und in vorbereiteten Texten eine Reihe von Phänomenen automatisch zu identifizieren. Die Teilnehmerzahl ist auf 30 beschränkt.

Lernziele und Methodik

Der Workshop verfolgt mehrere Ziele: (1) Er soll die Teilnehmer*innen mit spaCy und dessen Kernkomponenten vertraut machen und Ihnen praktische Erfahrung in der Nutzung von MONAPipe für typische Textanalysekomponenten auf Token-, Satz-/Teilsatz- und Passagenebene vermitteln. (2) Darüber hinaus erproben die Teilnehmer*innen die Einbindung neuer Komponenten, um damit wie sie MONAPipe für eigene Zwecke anpassen können. Aufbauend auf diesen Grundlagen lernen die Teilnehmer*innen an einem konkreten Beispiel, (3) wie sie MONAPipe konkret für Forschungsprojekte insbesondere in der digitales Literaturanalyse nutzen können. Dies umfasst die Auswahl geeigneter Komponenten für die Forschungsfrage sowie die Reflektion der Ergebnisse. Am Ende des Workshops haben die Teilnehmer*innen zum einen (i) ein besseres theoretisches Verständnis für die verschiedenen Sprachanalyseschritte, können komplexe Analysen durch Kombination mehrerer basaler Werkzeuge durchführen und die Qualität der automatischen Analyse beurteilen; Zum anderen (ii) haben die Teilnehmer*innen praktische Erfahrung im Umgang mit spaCy und verschieden Sprachverarbeitungswerkzeugen erworben und Problemlösungsstrategien für den Umgang mit NLP-Werkzeugen gelernt.

Methodisch kombiniert der Workshop Theorie und Praxis, wobei der Praxisanteil überwiegt. Um das Gelernte zu festigen und zu vertiefen, bekommen die Teilnehmer*innen zunächst kurze Arbeitsaufträge (zu den Sprachverarbeitungskomponenten) und später komplexere Aufgaben (zur Analyse narrativer Texte), deren Lösungen im Anschluss diskutiert werden. Der Praxisteil im zweiten Teil des Workshops bietet außerdem die Möglichkeit, MONAPipe für ein eigenes Forschungsproblem anzuwenden und dazu Feedback von den Organisator*innen des Workshops zu bekommen.

Auf technischer Ebene arbeiten wir mit der interaktiven Programmierumgebung Jupyter-Notebook und stellen vorbereitete und ausführlich dokumentierte Notebooks zur Verfügung, um einen möglichst reibungslosen Ablauf zu ermöglichen und den Teilnehmer*innen zu helfen, sich auf die Workshopinhalte zu konzentrieren.

Organisation und Ablauf

Wir planen einen vierstündigen Workshop bestehend aus zwei Blöcken. Der erste Block (1:45 h) beinhaltet aus einem einführenden Vortrag sowie einem Zeitslot zur Einrichtung der Jupyter-Notebooks , wobei die Organisator*innen nach Bedarf Hilfestellung bei der Einrichtung leisten. Anschließend erfolgt eine 45-minütige Session mit vorbereiteten Notebooks, bei der zunächst kürzere textuelle Phänomene auf Token-Ebene (wie Named Entities), Phänomene auf Teilsatz-Ebene (z. B. Zeitformen) sowie Phänomene, die längere Textpassagen umfassen (z. B. Redeformen), behandelt werden.

Im zweiten Block des Workshops (1:45 h) erstellen die Teilnehmer*innen eine eigene Komponente in spaCy. Anschließend erhalten die Teilnehmer*innen die Möglichkeit durch Lektüre narrative Strukturen in exemplarischen Textpassagen qualitativ zu bestimmen. Anhand der zur Verfügung stehenden spaCy-Komponenten soll evaluiert werden, welche Features sich zur Identifikation komplexer narrativer Strukturen eignen. Alternativ können die Teilnehmer*innen an eigenen Texten und Fragenstellungen arbeiten und hierfür Unterstützung durch die Workshoporganisator*innen erhalten.

Alle Teilnehmer*innen erhalten einen Gastaccount bei der GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen), um die Jupyter-Notebooks auf den GWDG-Jupyter-HPC-Servern nutzen zu können. Eine vorherige lokale Installation von Python oder zugehörigen Paketen ist nicht notwendig, dies wird im Vorfeld von der GWDG erledigt. Die HPC (High Performance Computing) Umgebung bietet die Möglichkeit, auch rechenaufwendige Pipelines zu testen. Im Rahmen der text- und sprachbasierten Forschungsdateninfrastruktur Text+ stellt die SUB Göttingen Schnittstellen bereit, mit denen literarische Texte aus der Digitalen Bibliothek in TextGrid direkt verwendet werden können.

Tabelle 1

Phase	Inhalt(e)	Zeit in Minuten
1. Einführung (Vortrag)	Grundkonzepte der maschinellen Sprachverarbeitung, narrativen Konzepten und der Programmbibliothek spaCy	20
2. Einrichtung Jupyter-Notebooks	Technische Einrichtung und Kurzüberblick zur Funktionsweise von Jupyter-Notebooks	20
3. Textuelle Phänomene (Vortrag + hands-on)	Vorbereitete Jupyter-Notebooks mit Aufgaben zu textuellen Phänomenen mit unterschiedlichen Spans: • Token-Ebene (Named Entities, Zeitmarker) • Teilsatzebene (Zeitformen) • Passagen (Redeformen)	1:05
Pause		25
4. Einbindung einer Custom-Komponente (hands-on)	Teilnehmer*innen integrieren eine eigene spaCy-Custom-Komponente (z. B. Fremdworterkennung)	45
5. Narrative Strukturen (hands-on + Diskussion)	Arbeitsaufgabe: narrative Strukturen in Texten erkennen	45
6. Abschluss		5

Nach dem Workshop

Wir tragen der Nachhaltigkeit der Forschung bei und stellen MONAPipe in einem Git-Repository zur Verfügung. Jupyter-Notebooks, die im Workshop benutzt werden, werden in einem separaten Git-Repository zur Verfügung gestellt.

Forschungsinteressen der Beitragenden

Hanna Varachkina , M. A., ist wissenschaftliche Mitarbeiterin und Doktorandin am Seminar für Deutsche Philologie der Universität Göttingen. Ihre Forschungsinteressen liegen in computergestützter Textanalyse: Modellierung und Erkennung von Textstrukturen und Diskurs-Phänomenen.

Florian Barth , M. A., ist wissenschaftlicher Mitarbeiter und Doktorand am Göttingen Centre for Digital Humanities und Mitarbeiter der Abteilung Forschung und Entwicklung der SUB Göttingen. Seine Forschungsinteressen liegen im Bereich der computationellen Textanalyse mit besonderem Fokus auf narrativen und fiktionstheoretischen Phänomenen sowie in der konkreten Anwendung dieser Forschung im Bereich der Infrastrukturen für die Digital Humanities.

Tillmann Dönicke , M. Sc., ist wissenschaftlicher Mitarbeiter und Doktorand am Göttingen Centre for Digital Humanities der Universität Göttingen. Seine Forschungsinteressen liegen in der strukturellen Textanalyse, insbesondere im Zusammenhang mit Narration und Fiktion, sowie der automatischen Erkennung narrativer Phänomene.

Johannes Biermann, M. A., ist wissenschaftlicher Mitarbeiter bei der Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG) im Bereich High Performance Computing (HPC). Die GWDG erfüllt u.a. die Funktion eines Rechen- und IT-Kompetenzzentrums für die Universität Göttingen. Im Zuge des Verbund für Nationales Hochleistungsrechnen (NHR-Verbund) ist er Berater für Anwendungen aus dem Bereich Digital Humanities. Sein Forschungsinteresse ist es, DH Fragestellungen auf High-Performance-Computing-Cluster zu adaptieren und dort zu rechnen.

Caroline Sporleder, ist Professorin für Digital Humanities am Institut für Informatik der Universität Göttingen und Leiterin des Göttingen Centre for Digital Humanities. Ihre Forschungsinteressen liegen im Bereich der computationellen Semantik und Diskursanalyse, besonders für Anwendungen der Geistes- und Kulturwissenschaften.

Bibliographie

Barth, Florian, Hanna Varachkina, Tillmann Dönicke, und Luisa Gödeke. 2022. "Levels of Non-Fictionality in Fictional Texts." In Proceedings of The Eighteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation . 27-32.

Brunner, Annelen, Ngoc Duyen Tanja Tu, Lukas Weimer, und Fotis Jannidis. 2020. "To BERT or not to BERT – comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing

representation." In 5th SwissText & 16th KONVENS Joint Conference 2020 .

Dönicke, Tillmann. 2020. "Clause-level tense, mood, voice and modality tagging for German." In Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories , 1–17.

Dönicke, Tillmann, Luisa Gödeke, und Hanna Varachkina. 2021. " Annotating Quantified Phenomena in Complex Sentence Structures Using the Example of Generalising Statements in Literary Texts ." In Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation , 20–32.

Dönicke, Tillmann, Florian Barth, Hanna Varachkina und andere. 2022. MONAPipe: Modes of Narration and Attribution Pipeline. (Softwarepublikation) URL: https://gitlab.gwdg.de/mona/pipy-public .

Gödeke, Luisa, Florian Barth, Tillmann Dönicke, Anna Mareike Weimer, Hanna Varachkina, Benjamin Gittel, Anke Holler und Caroline Sporleder. 2022 (zur Publikation angenommen). "Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung." In Zeitschrift für digitale Geisteswissenschaften .

Mohammad, Saif und Peter Turney. 2013. "Crowdsourcing a Word-Emotion Association Lexicon." In Computational Intelligence, 29 (3): 436-465.

Remus, Robert, Uwe Quasthoff, und Gerhard Heyer. 2010. "SentiWS - a publicly available German language resource for sentiment analysis." In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA). 1168-1171.

Tuggener, Don und Manfred Klenner. 2014. "A hybrid entity-mention pronoun resolution model for German using Markov logic networks . " In Proceedings of the 12th edition of the KONVENS conference, 21-31.

Vauth, Michael, Hans Ole Hatzel, Evelyn Gius and Chris Biemann. 2021. "Automated Event Annotation in Literary Texts." In CHR 2021: Computational Humanities Research Conference , November 17–19, 2021, Amsterdam, The Netherlands, 333-345.

Weimer, Anna Mareike, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Hanna Varachkina, Anke Holler, Caroline Sporleder und Benjamin Gittel. 2022 (zur Publikation angenommen). "The (In-)Consistency of Literary Concepts Operationalising, Annotating and Detecting Literary Comment." In Journal of Computational Literary Studies .