SANTA: Systematische Analyse Narrativer Texte durch Annotation

Gius, Evelyn
evelyn.gius@uni-hamburg.de
Universität Hamburg, Deutschland

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Strötgen, Jannik
jannik.stroetgen@mpi-inf.mpg.de
Max-Planck-Institut für Informatik, Saarbrücken, Deutschland

Willand, Marcus
marcus.willand@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

In diesem Beitrag wollen wir ein Vorhaben zur Diskussion stellen, das an zwei zentralen Herausforderungen in den Digital Humanities ansetzt: Der Erstellung adäquater Annotationsrichtlinien für geisteswissenschaftlich relevante textuelle Konzepte und der Schnittstelle in der Kooperation zwischen beteiligten Wissenschaftlerinnen und Wissenschaftlern aus Geisteswissenschaft und Informatik. Für DH-Projekte sind Kooperationen unerlässlich, wenn fortgeschrittene Techniken zur Textanalyse eingesetzt werden und/oder es um eine Zusammenführung von Konzepten oder Zugangsweisen geht, die bereits intradisziplinär als komplex gelten. Dabei wird ein signifikanter Anteil der Projektlaufzeit auf die Entwicklung einer "gemeinsamen Sprache" und die Identifikation der exakten, gemeinsamen wissenschaftlichen Fragestellung verwendet. Dies ist zweifellos ein produktiver Prozess, dessen erfolgreiche Durchführung allerdings voraussetzt, dass auf beiden Seiten Forscherinnen und Forscher beteiligt sind, die sich auf das interdisziplinäre Vorgehen voll einlassen und auch den nötigen Zeitaufwand tragen.

Methodisch-technisch ist ein substanzielles Nadelöhr bei der Entwicklung automatischer Werkzeuge das Fehlen von annotierten Goldstandards, an/auf denen Werkzeuge trainiert, verglichen und feinjustiert werden können. Das Fehlen der Goldstandards ist jedoch eigentlich ein nachgelagertes Problem, wie sich z.B. in narratologisch orientierten Projekten zeigt (heureCLÉA: Bögel et al., 2015; Propp annotation: Fisseni et al., 2014): Die Umsetzung narratologischer Theorien als Annotationen ist alles andere als trivial, da narratologische Konzepte nicht im Hinblick auf Annotation entwickelt wurden. Leerstellen in den Definitionen müssen gefüllt, Voraussetzungen geklärt und Unterkategorien geklärt werden. Die Annotation solcher Kategorien ist also kein reiner Umsetzungs- oder Implementierungsprozess, sondern einer bei dem sich tiefe, konzeptionelle Fragen stellen. Als Ergebnis solcher Prozesse stehen dann Annotationsrichtlinien, die die Brücke zwischen Theorie und Praxis schlagen. Erst wenn Annotationsrichtlinien für ein Phänomen (oder eine Gruppe von Phänomenen) etabliert sind, können größere Annotationsprojekte mit Aussicht auf Erfolg durchgeführt werden.

Das von uns vorgeschlagene Vorgehen erlaubt den Beteiligten Forscherinnen und Forschern ihre Expertise einzubringen, ohne in einem gemeinsamen Projektkontext zu arbeiten. Die Schnittstelle zwischen D und H wird hierbei von annotierten Daten und Annotationsrichtlinien gebildet, wobei die Richtlinien ohne Kompromisse bezüglich möglicher Automatisierungen erstellt werden. Das Vorhaben gibt somit auch narratologisch/literaturwissenschaftlich anspruchsvoller Konzeptentwicklung und damit Theoriebildung einen Rahmen. Verfügbare annotierte Daten wiederum erlauben Informatikerinnen und Informatikern ohne Expertise in narratologischen Fragen die Entwicklung von Werkzeugen für komplexe technische Probleme.

Ein *shared task* zur Erstellung von Annotationsrichtlinien

Shared Tasks sind in der Computerlinguistik weit verbreitet und haben für viele Bereiche gezeigt, dass sie ein geeignetes Instrument sind, Forschungsbemühungen verschiedener Gruppen zum gleichen Thema zu bündeln und zu verstärken. In einem *shared task* versuchen verschiedene Arbeitsgruppen mit verschiedenen Methoden dieselbe, klar definierte Aufgabe zu lösen, z.B. Word Sense Disambiguation (z.B. Mihalcea et al., 2004), Sentiment Analysis (z.B. Nakov et al., 2013) oder Named Entity Recognition (z.B. Sang and/De Meulder 2003). Auch wenn bisweilen im Rahmen von NLP-shared tasks Annotationsstandards neu entwickelt werden, liegt der Fokus hier auf der Verbesserung der Vorhersagequalität automatischer Systeme. Damit in einem solchen Vorgehen literaturwissenschaftlich relevante und interessante Konzepte und Phänomene bearbeitet werden, muss literaturwissenschaftliche Expertise bei der Erstellung der Annotationsrichtlinien einfließen.

Als Rahmen für die Entwicklung von Annotationsrichtlinien organisieren wir einen shared task der sich genau auf dieses Ziel konzentriert (Phase 1: Erstellung von Guidelines). Sind die Richtlinien etabliert, kann anschließend ein großes Korpus annotiert werden,

das wiederum in einem NLP-shared task eingesetzt werden kann, um Verfahren zu erproben, die die annotierten Phänomene automatisch finden (Phase 2: Automatisierung).

Als Phänomen haben wir uns dabei auf Erzählebenen in englischen und deutschen Texten festgelegt, da diese für zahlreiche, komplexere literaturwissenschaftliche Fragestellungen hilfreich sind, ohne selbst (für einen ersten *shared task*) zu komplex zu sein. Zudem sind sie als Phänomen omnipräsent: Praktisch jeder narrative Text enthält mehr als eine Erzählebene, und sie sind auch in nicht-textuellen Medien wie z.B. Filmen verbreitet. Die Existenz verschiedener Theorien zur Analyse von Erzählebenen in literarischen Texten belegt, dass es dabei auch konzeptionellen, theoretischen Entwicklungsbedarf gibt. Erzählebenen bilden darüber hinaus eine wichtige Segmentierungsstufe für die weitere automatische semantische Verarbeitung von Texten: z.B. sollte Koreferenzresolution von der vorher erfolgten Erkennung von Erzählebenen profitieren,da Koreferenzketten in heterodiegetischen eingebetteten Erzählungen nicht ebenenübergreifend sein sollten.

Während Details zum Gesamtaufbau des Shared Tasks bereits in einem anderen Artikel beschrieben wurden (Reiter et al., 2017), fokussieren wir uns in diesem Beitrag auf die genauere Beschreibung der ersten Phase des *shared tasks*.

Geplanter Ablauf

Erstellung von Annotationsrichtlinien

(bis Mitte Juni 2018)

Im ersten Schritt wird allen Teilnehmerinnen und Teilnehmern ein *development corpus* bestehend aus ca. 20 Texten zugänglich gemacht. Die Texte liegen auf deutsch und englisch vor und decken verschiedene Genres und Epochen ab. Die Texte enthalten verschiedene Arten von Erzählebenen, gemäß eines etwas vagen Vorverständnisses.

Die Texte können und sollen von den Teilnehmerinnen und Teilnehmern benutzt werden, um Richtlinien für die Annotation von Erzählebenen zu entwickeln und zu testen. Ob die Texte in einer oder in beiden Sprachen verwendet werden, ist dabei den Teilnehmerinnen und Teilnehmern überlassen. Sie sollten dabei das Ziel verfolgen, eine möglichst breite Anwendbarkeit der Richtlinien sicherzustellen (auch jenseits des *development corpus*). Außerdem sollen die Richtlinien vollständig und selbsterklärend sein, so dass kein Expertenwissen zur Anwendung vorausgesetzt wird. Um mehrsprachige Anwendung zu ermöglichen, sollen die Richtlinien auf Englisch formuliert sein, sie dürfen aber sprachspezifische Beispiele enthalten.

Wie genau die Gruppen dabei vorgehen, bleibt ihnen überlassen. In vergangenen Annotationsprojekten (mit und ohne Bezug zu Literaturwissenschaft bzw. literarischen Texten) hat sich aber ein iterativer Prozess als fruchtbar erwiesen. Sobald eine erste Version der Richtlinien erstellt wurde, werden sie auf neuen Texten getestet, um ihre Definitionslücken oder Vagheiten zu identifizieren. Aus dem Schließen der Lücken ergibt sich dann eine weitere Version der Richtlinien, die wiederum auf Texten getestet werden können.

Anwendung eigener Guidelines

(bis Ende Juni 2018)

Im zweiten Schritt sollen die Arbeitsgruppen ihre eigenen Richtlinien auf neuen Texten testen. Nach dem Einreichen ihrer Richtlinien erhalten die Teilnehmerinnen und Teilnehmer hierzu sechs neue literarische Texte, die vom Organisationsteams des *shared tasks* ausgesucht wurden. Die Annotation dieser Texte muss dabei in einem Web-basierten, frei zugänglichen, von den Organisatoren bereitgestellten Annotationstool durchgeführt werden, um die automatisierte Auswertung der Annotationen und ihren Vergleich zu ermöglichen.

Anwendung von Guidelines anderer Teilnehmer

(bis Mitte Juli 2018)

Im dritten Schritt erhält jede teilnehmende Gruppe Richtlinien anderer Gruppen, auf deren Basis Erzählebenen in den sechs Texten erneut annotiert werden, wobei alle Richtlinien von uns zuvor anonymisiert werden. Zusätzlich wird auch eine vom Organisationsteam betreute Gruppe von studentischen Hilfskräften alle eingereichten Annotationsrichtlinien auf den sechs Texten anwenden.

Evaluation aller vorgeschlagener Guidelines

(August/September 2018)

Im letzten Schritt der ersten Phase des *shared tasks* werden alle eingereichten Annotationsrichtlinien verglichen und evaluiert. Dafür treffen sich die Teilnehmerinnen und Teilnehmer zu einem Workshop, auf dem sie ihre eigenen Richtlinien vorstellen und gemeinsam Qualität und Komplexität bewertet werden. Das Ziel des Workshops ist außerdem, basierend auf der Diskussion und den Informationen bezüglich der Inter-Annotator-Agreements im Plenum und möglichst konsensual die Annotationsrichtlinien zu bestimmen, die dann in der zweiten Phase des *shared tasks* verwendet werden. Auf deren Basis werden dann Methoden und Systeme entwickelt, die automatisch Erzählebenen in Texten identifizieren können.

Zur vergleichenden Evaluation von Annotationsrichtlinien sind bisher Ansätze aus der Computer- und Korpuslinguistik zur quantitativen Messung des Inter-Annotator-Agreement (IAA) bekannt (vgl. Artstein, 2017), die im Bereich der Digital Humanities

angewendet wurden und werden. Da es aber bei der Erstellung von Annotationsrichtlinien für narratologische Phänomene eben nicht *nur* um die Umsetzung und Erklärung einer klar spezifizierten Theorie geht, sondern eben *auch* um die (Weiter-)Entwicklung narratologischer Konzepte, bedarf es eines weitergehenden Blickes. Dabei sollen drei Aspekte Berücksichtigung finden: Die **Anwendbarkeit** von Annotationsrichtlinien kann durch quantitatives IAA gemessen werden. Hier stellen sich durch möglicherweise unterschiedliche theoretische Zugänge vor allem Fragen der Vergleichbarkeit. Der Aspekt der begrifflichen **Abdeckung** bezieht sich darauf, welche (bekannten) narratologischen Ebenenkonzeptionen in der konkreten Ausgestaltung vollständig oder teilweise enthalten sind. Dies wird sich nur durch qualitative Analyse und wissenschaftliche Diskussion basierend auf theoretischen Vorstudien klären lassen, für die der Workshop einen Rahmen bieten soll. Die **Nützlichkeit** von Annotationsrichtlinien kann bei narrativen Ebenen dahingehend bewertet werden, ob sie interpretativ wertvolle Beschreibungen erlauben. Leitgedanke ist hier, dass narratologische Annotationen eine deskriptive Basis für literaturwissenschaftliche Interpretationen liefern sollen. Unterschiedlichen Annotationsrichtlinien zu folgen hieße also zu unterschiedlichen Text-Deskriptionen zu kommen, die wiederum unterschiedliche Interpretationen zulassen.

Conclusions

Im Rahmen des Vortrags wollen wir insbesondere zwei der o.g. Aspekte in den Fokus rücken und diskutieren: Die iterative Entwicklung von Annotationsrichtlinien als verteiltes, kollaboratives Projekt sowie die Evaluation und Vergleichbarkeit von Annotationsrichtlinien für literarische Phänomene.

Bibliographie

Artstein, Ron (2017): “Inter-annotator Agreement”, in: Ide, Nancy / Pustejovsky James (eds.): *Handbook of Linguistic Annotation*. Dordrecht: Springer. DOI 10.1007/978-94-024-0881-2.

Bögel, Thomas / Gertz, Michael / Gius, Evelyn / Jacke, Janina / Meister, Jan Christoph / Petris, Marco / Strötgen, Jannik (2015): “Collaborative text annotation meets machine learning: heurecléa, a digital heuristic of narrative”, in: DHCommons 1.

Fisseni, Bernhard / Kurji, Aadil / Löwe, Benedikt (2014): “Annotating with Propp’s morphology of the folktale: Reproducibility and trainability”, in: *Literary and Linguistic Computing* 29(4):488–510, 1093/llc/fqu050

Mihalcea, Rada / Chklovski, Timothy / Kilgarriff, Adam (2004): “The Senseval-3 English Lexical Sample Task”. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

Nakov, Preslav / Rosenthal, Sara / Kozareva, Zornitsa / Stoyanov, Veselin / Ritter, Alan / Wilson, Theresa (2013): “SemEval-2013 Task 2: Sentiment Analysis in Twitter”. In *Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval ‘13*, Atlanta, Georgia, USA.

Reiter, Nils / Gius, Evelyn / Strötgen, Jannik / Willand, Marcus (2017): “A Shared Task for a Shared Goal - Systematic Annotation of Literary Texts ”. In *Digital Humanities 2017: Conference Abstracts*, Montreal, Canada.

Sang, Erik F. Tjong Kim / de Meulder, Fien (2003): “Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition”, in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*.