CUTE: CRETA Unshared Task zu Entitätenreferenzen

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

Blessing, Andre
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Echelmeyer, Nora
nora.echelmeyer@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Koch, Steffen
steffen.koch@vis.uni-stuttgart.de
Universität Stuttgart, Deutschland

Kremer, Gerhard
gerhard.kremer@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Murr, Sandra
sandra.murr@ilw.uni-stuttgart.de
Universität Stuttgart, Deutschland

Overbeck, Maximilian
maximilian.overbeck@sowi.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pichler, Axel
axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart, Deutschland

Einleitung

Der Workshop zum CRETA Unshared Task (CUTE) verfolgt ein inhaltliches und ein methodisches Ziel. Das inhaltliche Ziel ist die Anregung eines Diskurses über Entitäten, deren Annotation und Kategorisierung entlang von geistes- und sozialwissenschaftlichen Forschungsfragen sowie deren Potential als disziplinübergreifende Textanalyseaufgabe. Methodisch möchten wir ein Workshop-Format erproben, das unseres Erachtens eine produktive Schnittstelle zwischen Geistes-/SozialwissenschaftlerInnen und InformatikerInnen bildet. Das genaue Programm des Workshops wird von den Teilnehmenden durch Beiträge gestaltet (durch Beiträge, siehe Call for Papers) und vor rechtzeitig vor dem Workshop auf der Webseite veröffentlicht .

Entitätenreferenzen

Das Konzept der Entität und ihrer Referenz ist ein bewusst weites, das anschlussfähig sein soll für verschiedene Forschungsfragen aus den Geistes- und Sozialwissenschaften. Wir möchten dabei explizit verschiedene Perspektiven auf Entitäten berücksichtigen.

Entitäten in der Literaturwissenschaft

Figuren in literarischen Texten sind „mit ihrer sinnkonstitutiven und handlungsprogressiven Funktion“ ein zentraler Bestandteil der fiktiven Welt (Platz-Waury 1997). Von besonderem Interesse dabei sind Figurenkonstellationen und Interaktionen, die Entwicklung von Figuren sowie die Funktionalisierung von Figuren als Handlungsträger. Die Erkennung von Figurenreferenzen ist grundlegend, um z.B. Figuren zu charakterisieren, ihre Relationen identifizieren und Netzwerkanalysen durchführen zu können (vgl. Jannidis 2015, Trilcke 2013).

Neben der Figur rückt –- spätestens seit dem *spatial turn* –- auch der Raum als relevante Entität in den Fokus der Literaturwissenschaft. Der Handlungsraum in literarischen Texten dient der Strukturierung der fiktiven Welt und ist zumeist semantisiert (Lotman 1972). Zudem kann er in Wechselwirkung mit Aspekten der Figur („sujethafte Grenzüberschreitung“, Lotman 1972) oder der Zeit stehen („Chronotopos“, Bachtin 1989).

Entitäten in der Sozialwissenschaft

Politische Parteien, internationale Organisationen oder Institutionen sind seit jeher zentrale Analyseobjekte der empirischen sozialwissenschaftlichen Forschung und werden spätestens seit dem *linguistic turn* (Rorty 1967) in den Sozialwissenschaften auch mittels inhalts- oder diskursanalytischer Methoden auf zunächst kleinen und zunehmend größeren Mengen von Textdokumenten (beispielsweise Parteiprogrammen, offizielle Regierungsdokumenten, Zeitungstexten) untersucht. Neben vielfältigen anderen Analysen stehen dabei oftmals Fragen nach der Sichtbarkeit oder Bewertung bestimmter Entitäten, wie beispielsweise der Europäischen Union als supra-/internationaler Organisation (Kantner 2015) im Vordergrund.

Entitäten in der Philosophie

Im Unterschied zu den Literatur- und Sozialwissenschaften spielen Entitäten als Untersuchungsgegenstand in philosophischen Texten zunächst keine Rolle. Aufgrund ihrer metareflexiven

Ausrichtung fragt Philosophie primär nicht nach individuell unterscheidbaren Objekten in der echten oder einer fiktiven Welt, sondern beschäftigt sich mit transzendentalen Fragen nach den Bedingungen und Möglichkeiten derartiger individueller Objekte. Dabei arbeitet sie mit abstrakten Konzepten, die sich ebenfalls als -- nicht-dingliche -- Objekte einer Welt auffassen lassen. Pragmatisch gesehen erfolgt die Referenz auf abstrakte Konzepte in Texten jedenfalls in ähnlicher Weise wie die Referenz auf Figuren, Organisationen und Orten (s.u.).

Fachübergreifende Annotationsschemata

Auch wenn die Interpretation von z.B. der Erwähnung von Organisationen in politischen und des Auftretens von Figuren in literarischen Texten anderen Regeln folgt und mit anderen Forschungsfragen zusammenhängt, gibt es Gemeinsamkeiten auf linguistisch-struktureller Ebene. Im Text realisiert werden Referenzen auf die o.g. Arten von Entitäten entweder als Eigennamen (*Angela Merkel*/ *Ästhetische Theorie*), Pronomen (*sie*/ *sie*) oder als appellative Nominalphrasen (*die Bundeskanzlerin*/ *das Spätwerk Adornos*). Wir haben daher ein einheitliches Vokabular und Annotationsschema entwickelt und auf einem ausgewählten heterogenen Korpus getestet. Dieses soll im Rahmen des Workshops diskursiv erörtert und wenn möglich erweitert werden.

Abstrakt gesprochen verstehen wir unter Entitäten individuell unterscheidbare Objekte in der echten oder einer fiktiven Welt. Wir unterscheiden sechs verschiedene Typen von Entität: Personen, Orte, Ereignisse, Organisationen, kulturelle Artefakte und Konzepte. Die Bezeichnung als „Objekt“ impliziert also *nicht*, dass es sich um physikalische Objekte handelt. Die Einteilung in Typen ist von den oben skizzierten Forschungsfragen und -feldern abgeleitet und ist -- bei anderen Forschungsfragen oder -daten -- offen für Ergänzungen. Die Anwendbarkeit auf zusätzliche Texte und Textgattungen ist für uns (und für diesen Workshop) von besonderem Interesse.

Die Erstellung abstrakter Annotationsrichtlinien und deren systematische, kontrollierte Anwendung (Annotation) auf konkrete Texte verspricht im Wesentlichen zwei Ergebnisse:

1. Das Erzeugen von parallelen Annotationen auf Basis von Richtlinien zwingt zu einem sehr genauen Lesen des Textes und sorgt für eine intensive Auseinandersetzung mit den Annotationskategorien (und auch für ein Hinterfragen derselben). Recht schnell wird auf diese Weise deutlich, welche Annahmen bei der Anfertigung der Annotationsrichtlinien nicht von den Daten gedeckt waren. Auch Phänomene, die inhaltlich berücksichtigt werden sollten, aber nicht in den Richtlinien enthalten sind, fallen den FachwissenschaftlerInnen schnell ins Auge. Dadurch, dass die eigenen Annotationsentscheidungen ggf. diskutiert und verteidigt werden müssen, sorgen Parallelannotationen für die Aufdeckung von Vagheiten in den Definitionen und damit für eine Klärung der Begriffe (vgl. Gius / Jacke 2016).
2. Die Entwicklung von maßgeschneiderten Textanalysewerkzeugen für spezifische geistes- und sozialwissenschaftliche Forschungsfragen stößt schnell an Ressourcengrenzen. Als Problem erweist sich oft, dass die Textanalyseaufgaben zu speziell oder die Datenmengen zu klein sind und damit ein Forschungsbeitrag in der Informatik oder Computerlinguistik nur schwer möglich ist (was typischerweise wiederum Auswirkungen auf den Ressourceneinsatz hat). Eine Antwort auf diese Herausforderung ist die Etablierung fachübergreifender Textanalyseaufgaben, etwa für bestimmte Annotationsebenen. Dies erlaubt die Entwicklung von allgemeineren, wiederverwendbaren Werkzeugen und -- mit geeigneten Testdaten -- deren iterative Verbesserung. Damit wird die Bearbeitung geistes- und sozialwissenschaftlicher Forschungsfragen letztlich nachhaltiger unterstützt als durch die Entwicklung spezieller, aber nach Projektende nicht weiterentwickelter Werkzeuge. Ein Katalysator dafür können shared und unshared tasks sein (vgl. Kuhn / Reiter 2015).

Shared/Unshared Task

In diesem Sinne ist das zweite, methodische Ziel des Workshops zu verstehen: Wir möchten einen Community-Task veranstalten, der eine *shared* und drei *unshared*-Tracks hat. Damit wird ein Workshop-Format auf die Probe gestellt, das eine produktive Schnittstelle zwischen Geistes-/ SozialwissenschaftlerInnen und InformatikerInnen zu bilden verspricht (s.a. Belz / Kilgarriff 2006). Im Gegensatz zu *shared tasks*, bei denen die Performanz verschiedener Systeme, Ansätze oder Methoden direkt anhand einer klar definierten und quantitativ evaluierten Aufgabe verglichen wird, sind *unshared tasks* offen für verschiedenartige Beiträge, die auf einer gemeinsamen Datengrundlage oder Fragestellung basieren. Neben dem Call -- der bereits eine Sammlung möglicher Fragestellungen nennt -- veröffentlichen wir daher ein heterogenes Korpus, das als Datengrundlage dient. Im Rahmen von CUTE können Forscherinnen und Forscher an den folgenden Tracks teilnehmen:

1. **Automatische Erkennung von Entitätenreferenzen**: Experimente zum automatischen Vorhersagen von Annotationen auf noch nicht annotierten Texten, mit regelbasierten oder statistischen Systemen

2. **Visualisieren von Entitätenreferenzen im Text**: Visualisierungsmöglichkeiten zur (interaktiven) Exploration der vorhandenen oder neuen Annotationen
3. **Annotationsanalyse**: Qualitative oder quantitative Analyse der vorhandenen Annotationen oder der Annotationsrichtlinien; Annotationsexperimente zur Anwendbarkeit der Richtlinien auf neue Texte
4. **Freestyle**: Kreative Ideen, die keinen der obigen Tasks adressieren

Beiträge zu Aufgabe 1 werden quantitativ evaluiert und im Wettbewerb mit den Evaluationsergebnissen der anderen Beiträge verglichen (*shared task*, die technischen Details dazu werden auf der Webseite veröffentlicht). Beiträge für die Aufgaben 2 bis 4 werden vom Programmkomitee qualitativ evaluiert (*unshared task*). Der Austausch während des Workshops (in Form von Kurzvorträgen und Diskussion) wird insoweit eine Bandbreite an Zugängen abbilden, deren verbindendes Element die gemeinsame Datengrundlage sein wird. Da die Teilnehmerinnen und Teilnehmer sich dann im Vorfeld intensiv mit den Daten aus verschiedenen Perspektiven beschäftigen werden, erwarten wir für den Workshop eine erkenntnisreiche Diskussion.

Textgrundlage und Daten

Das von uns im Rahmen des Workshops veröffentlichte Korpus umfasst vier Teilkorpora:

1. jeweils eine PolitikerInnenrede aus insgesamt vier Parlamentsdebatten des Deutschen Bundestags (S. Leutheuser-Schnarrenberger am 28.10.99, A. Merkel am 16.12.04, A. Ulrich am 15.11.07 und A. Karl am 17.03.11)
2. Briefe aus Goethes *Die Leiden des jungen Werther* (1787) vom 4. Mai bis einschließlich 16. Juni
3. der Abschnitt Zur Theorie des Kunstwerks aus Adornos Ästhetische Theorie
4. die Bücher 3 bis 6 aus Wolframs von Eschenbach *Parzival* (mittelhochdeutsch)

Auch wenn jedes Teilkorpus seine eigenen Besonderheiten hat, wurden alle nach einheitlichen Annotationsrichtlinien annotiert, die wir ebenfalls veröffentlichen und zur Diskussion stellen möchten.

Ausrichter

Der Workshop wird ausgerichtet vom Centre for Reflected Text Analytics (CRETA) an der Universität Stuttgart. CRETA verbindet Literaturwissenschaft, Linguistik, Philosophie und Sozialwissenschaft mit Maschineller Sprachverarbeitung und Visualisierung. Hauptaufgabe von CRETA ist die Entwicklung reflektierter Methoden zur Textanalyse, wobei wir Methoden als Gesamtpaket aus konzeptuellem Rahmen, Annahmen, technischer Implementierung und Interpretationsanleitung verstehen. Methoden sollen also keine "black box" sein, sondern auch für nicht-Technikerinnen und -Techniker so transparent sein, dass ihr reflektierter Einsatz im Hinblick auf geistes- und sozialwissenschaftliche Fragestellungen möglich wird.

Fußnoten

1. http://dhd-blog.org/?p=7333
2. http://www.creta.uni-stuttgart.de/index.php/de/cute/
3. Von dem in der maschinellen Sprachverarbeitung etablierten Task der *named entity recognition* (NER) unterscheidet sich die vorliegende Aufgabe insofern, als dass unsere Annotationen neben Eigennamen auch andere Arten von Referenz enthalten. Werkzeuge (und tasks) zur NER sind darauf getrimmt, ausschließlich Eigennamen zu erkennen.

Bibliographie

Bachtin, Michail Michailowitsch / Kowalski, Edward / Wegner, Michael (1989): *Formen der Zeit im Roman. Untersuchungen zur historischen Poetik*. Frankfurt am Main: Fischer.

Belz, Anja / Kilgarriff, Adam (2006): „Shared-task Evaluations in HLT: Lessons for NLG“, in: *Proceedings of the Fourth International Natural Language Generation Conference*.

Gius, Evelyn / Jacke, Janina (2016): „Kollaboratives Annotieren literarischer Texte“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung*.

Jannidis, Fotis / Krug, Markus / Reger, Isabella / Toepfer, Martin / Weimer, Lukas / Puppe, Frank (2015): „Automatische Erkennung von Figuren in deutschsprachigen Romanen“, in: *DHd 2016: Von Daten zu Erkenntnissen*.

Kantner, Cathleen (2015): *War and Intervention in the Transnational Public Sphere: Problem-solving and European identity-formation*. New York: Routledge.

Kuhn, Jonas / Reiter, Nils (2015): „A Plea for a Method-Driven Agenda in the Digital Humanities“, in: *DH2015: Global Digital Humanities*.

Lotman, Juri (1972): *Die Struktur literarischer Texte*. München: Fink.

Platz-Waury, Elke (1997): „Figur“, in: Weimar, Klaus (ed.): *Reallexikon der deutschen Literaturwissenschaft*. Neubearbeitung des Reallexikon der deutschen Literaturgeschichte. Berlin, New York: de Gruyter 587–589.

Rorty, Richard M. (1967): *The Linguistic Turn*. Chicago: University of Chicago Press.

Trilcke, Peer (2013): „Social Network Analysis als Methode einer textempirischen Literaturwissenschaft“, in: Ajouri, Philip / Mellmann, Katja / Rauen, Christoph (eds.):

Empirie in der Literaturwissenschaft. Münster: Mentis 201–247.