Entitäten im Fokus am Beispiel von Captivity Narratives

Kessler, Linda
st150918@stud.uni-stuttgart.de
Universität Stuttgart, Deutschland

Braun, Tamara
st151509@stud.uni-stuttgart.de
Universität Stuttgart, Deutschland

Preuß, Tanja
st102459@stud.uni-stuttgart.de
Universität Stuttgart, Deutschland

Eigennamenerkennung (NER) ist im Bereich der maschinellen Sprachverarbeitung bereits viel behandelt worden. Eine Übersicht hierzu findet sich bei Nadeau und Sekine (2007). In den Digital Humanities dient die Erkennung von benannten Entitäten der Identifikation zentraler Akteure und Elemente in Texten, welche unter anderem die Grundlage für tiefergehende Analysen bezüglich Beziehungen, Strukturen und Emotionen in diesen Texten bilden. Jannidis et al. (2015) thematisieren allerdings, dass die reine NER beispielsweise für eine Analyse von Figurennetzwerken in literarischen Texten unzureichend ist, da dabei nur Figurenreferenzen durch konkrete Namensnennung erfasst werden. Um spezifisch auf die Bedürfnisse von Textanalysen im Kontext der Digital Humanities einzugehen, wurden im „Center for Reflected Text Analytics“ (CRETA) (Kuhn et al. 2016) der Universität Stuttgart Annotationsrichtlinien entworfen, die über die Annotation reiner Eigennamen hinausgehen und sich auf verschiedenartige Entitätsreferenzen in deutschsprachigen Texten unterschiedlicher Genres fokussieren.[1] So wird beispielsweise das Appellativ *the indians* als Entität erfasst, obwohl die Referenz nicht mit Namen spezifiziert wird.

Ein Beispiel für den Mehrwert der Annotation solcher Entitätsreferenzen findet sich bei Blessing et al. (2017). Um die Übertragbarkeit der Richtlinien nicht nur zwischen verschiedenen Textsorten, sondern auch sprachübergreifend zu evaluieren, stellen wir unser Projekt mit dem Ziel der Annotation von Erzähltexten in englischer Sprache vor. Ausgehend von der durch CRETA geschaffenen Grundlage teilt sich unser Projekt in drei Phasen auf: die manuelle Annotation und Überprüfung der Übertragbarkeit der CRETA-Richtlinien auf die gegebene Textsorte, die Automatisierung der Entitätserkennung und die Einbindung der Entitäten in eine literaturwissenschaftliche Analyse.

Als Textgrundlage dient eine Sammlung von englischsprachigen Captivity Narratives.[2] Diese Erzählungen aus dem 18. Jahrhundert handeln von Erfahrungen weißer Siedler in Nordamerika, die in indianische Gefangenschaft geraten. Zunächst wurden in sieben Texten im Gesamtumfang von 71.526 Wörtern 5.163 Entitäten identifiziert und mit den von CRETA erarbeiteten Kategorien (Personen, Orte, Organisationen, Ereignisse, Werke und abstrakte Konzepte) annotiert. Im Verlauf dieser Annotationsphase wurden die CRETA Richtlinien an die speziellen Gegebenheiten der Textsorte angepasst, die in Bezug auf die Erwähnung von Personen und Orten einige Besonderheiten aufweist. Auffällig ist beispielsweise, dass Personen in vielen Fällen in Gruppen, oftmals auch ohne spezifische Namen, erwähnt werden. Um diese Nennungen dennoch zu erfassen, wird die Begrenzung auf Eigen- und Gattungsnamen aufgehoben und um Formen wie *some, others* oder *a few* erweitert. Zudem werden Orte häufig anhand von Landschaftsmerkmalen und nur selten mit konkreten Ortsnamen benannt. Dementsprechend bilden solche Nennungen (z.B. *the river* oder *the mountain*) den Großteil der annotierten Ortsentitäten. Die Erfassung von vollständigen Nominalphrasen als Entitäten erweist sich stellenweise als problematisch, da die Captivity Narratives verschachtelte Nominalphrasen enthalten, sodass sehr umfangreiche Entitäten zu annotieren sind.

Der so entstandene Goldstandard dient als Trainingsdatensatz zur Entwicklung eines maschinellen Lernverfahrens. Ein Naive Bayes Classifier wurde mit Features trainiert, die sich u.a. auf die äußere Gestalt (z.B. Großschreibung), die Wortart und die Zugehörigkeit zu Wortlisten (Namen und amerikanische Orte) beziehen. Im Kreuzvalidierungsverfahren kann damit ein Micro-Fscore von 0,29 erzielt werden. Für die am häufigsten im Trainingsmaterial vorhandene Klasse PER wurde ein Precision-Wert von 0,45 erzielt. Dies bedeutet, dass fast die Hälfte der automatisch mit PER annotierten Entitäten wirklich Personen sind. Der Recall von 0,3 zeigt, wie unvollständig die Erkennung mit einem knappen Drittel aller relevanten Personen noch ist. Eine Auswertung der Ergebnisse zeigt, dass die Länge und Verschachtelung vieler Entitäten die automatische Klassifizierung erschwert. Da sich im manuellen Annotationsprozess der Kontext häufig als Entscheidungshilfe herausstellte, sollte dieser bei der automatischen NER zukünftig berücksichtigt werden. Darüber hinaus könnte die Erweiterung der verwendeten Features durch syntaktische Informationen und die Verwendung einer größeren Menge an Trainingsdaten zu Verbesserungen führen.

Um den Mehrwert der Entitätsreferenzen für eine inhaltliche Fragestellung bezüglich der Captivity Narratives zu veranschaulichen, zeigen wir die textstatistische Analyse von Emotionen im Umfeld bestimmter Entitäten bzw. Entitätsgruppen. Basierend auf den manuell annotierten Texten, lassen sich Personenentitäten mithilfe von Clusteranalysen gruppieren.

Anhand von positiven und negativen Wortlisten lassen sich zwei Gruppen bilden, die sich grob als *Indianer* und *Andere* gegenüber stehen (siehe Abbildung 1 und 2). Eine auf denselben Wortlisten basierende Sentiment Analyse ergab einen deutlich negativeren Emotionswert für Personenentitäten, die der Gruppe der Indianer zuzuordnen sind, als für die Gruppe der anderen Personen.

Abschließend lässt sich festhalten, dass auf Grundlage unserer Annotationen eine Abgrenzung der im Text auftretenden Gruppen anhand von emotionsgeladenen Wörtern möglich ist, die der erwarteten negativen Haltung der Verfasser gegenüber den Eingeborenen Nordamerikas entspricht.

Die von CRETA entwickelten Annotationsrichtlinien sind grundsätzlich auf die von uns analysierten Texte anwendbar, trotz abweichender Sprache und spezifischer Erzählweise. Um die Breite der enthaltenen Entitätsreferenzen vollständig abbilden zu können, bedarf es allerdings einzelner Spezifizierungen der Annotationsrichtlinien für diese Textsorte.

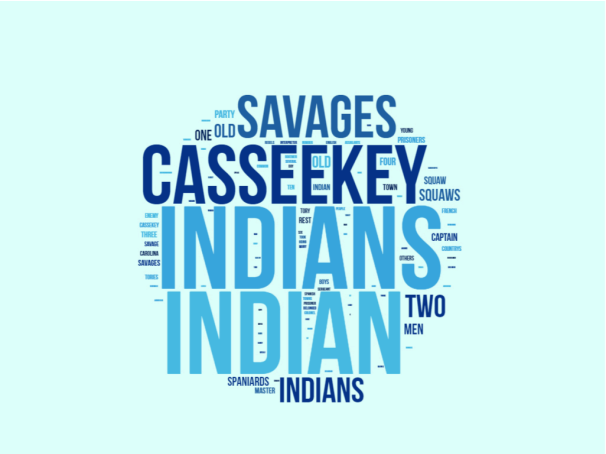

Abbildung 1: Ergebnisse der Cluster-Analyse: Indianer

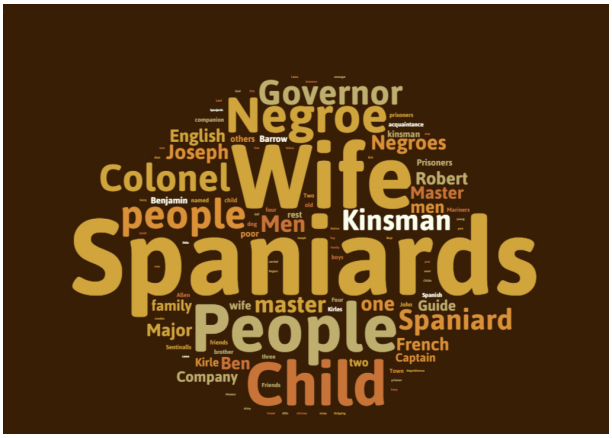

Abbildung 2: Ergebnisse der Cluster-Analyse: Andere

Fußnoten

1. Die Annotationsrichtlinien können eingesehen werden unter https://www.creta.uni-stuttgart.de/cute/datenmaterial/annotationsrichtlinien-1-1/ , abgerufen am 24.08.2017.
2. Wir danken Herrn Prof. Dr. Marc Priewe und den Mitarbeiter/innen der Professur für amerikanische Literatur und Kultur an der Universität Stuttgart

Bibliographie

Blessing, Andre / Echelmeyer, Nora / John, Markus / Reiter, Nils (2017): “An End-to-end Environment for Research Question-Driven Entity Extraction and Network Analysis”, in: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 57–67, Vancouver, Canada. Association for Computational Linguistics.

Jannidis, Fotis / Krug, Markus / Toepfer, Martin / Puppe, Frank / Reger, Isabelle / Weimer, Lukas (2015): “Automatische Erkennung von Figuren in deutschsprachigen Romanen“. Abstract für die DHd 2015 in Graz.

Kuhn, Jonas / Alexiadou, Artemis / Braun, Manuel / Ertl, Thomas / Holtz, Sabine / Kantner, Catleen, . . . Zittel, Claus (2016): “CRETA (Centrum für reflektierte Textanalyse) – Fachübergreifende Methodenentwicklung in den Digital Humanities“. Abstract für die DHd 2016 in Leipzig.

Nadeau, David / Sekine, Satoshi (2007): “A survey of named entity recognition and classification. *Linguisticae Investigationes”*, *30*, 3-26.