Annotation natürlichsprachlicher Texte aus Onlineforen zur Entwicklung domainspezifischer Ontologien

Hastik, Canan
hastik@linglit.tu-darmstadt.de
TU Darmstadt, Deutschland

Annotation natürlicher Sprachdaten aus sozialen Medien zur Erforschung zeitgenössischer Szenen, zur Sprach- und Trendanalyse und zur Weiterentwicklung von Sprachtechnologien gewinnt mit der zunehmenden Verfügbarkeit großer Datenbestände weiter an Bedeutung (Farzindar / Inkpen 2015). Zeitgenössische Kommunikation in sozialen Medien verfügt über inhaltliche und strukturelle Besonderheiten und ist von umgangssprachlicher Ausdrucksform geprägt. Beiträge, die im Kontext internetbasierter Diskussionskulturen in Foren entstehen, stellen eine wichtige Forschungsquelle dar. Diese nutzergenerierten Texte, in Form von semi- oder unstrukturierten Kommentaren, repräsentieren Meinungen und Bewertungen einer Gemeinschaft zu einem Thema, Produkt oder Werk und beziehen sich in der Regel auf inhaltliche, technische oder ästhetische Aspekte. Die Autoren verwenden dabei Sprachmittel wie Metaphern, Analogien, Ambiguität, Humor und Ironie sowie metalinguistische bildhafte Mittel wie Emoticons oder andere graphische Zeichen (Reyes et al. 2012).

Vor diesem Hintergrund adressiert dieses Projekt Herausforderungen, die bei der linguistischen und statistischen Verarbeitung von realen web-basierten Daten entstehen. Es wird ein Ansatz semi-automatischer Annotation zur Extraktion von Begriffen für die ontologiebasierte Beschreibung von computergenerierten audiovisuellen Kunstwerken einer digitalen Kunstszene präsentiert. Forschungsgegenstand ist die Diskussionskultur der Demoszene, einer spezialisierten Computerkunstszene. Bisher sind die zahlreichen Beiträge der Gemeinschaft, die sich auf ästhetische und technische Aspekte der Kunstwerke beziehen, nicht erschlossen. Bei diesen Beiträgen handelt es sich um informelle, emotionale, kurze und unstrukturierte Kommentartexte. Das verwendete Vokabular ist mehrsprachig und beinhaltet fachspezifische Terminologien, exklusive Neologismen und einen eigenen szenespezifischen orthographischen Stil. Diese Beiträge bieten detaillierte Einblicke in die Charakteristika der Werke, weshalb ihre Erschließung deren Verständnis fördert und eine gezielte Recherche einzelner Werke ermöglicht. Das Projekt befasst sich mit der Fragestellung, in wieweit sich aktuelle Verfahren der natürlichen Sprachverarbeitung (NLP), die auf grammatikalisch korrekte Schriftformen optimiert und auf Zeitungskorpora trainiert sind, anwenden lassen. Somit leistet das präsentierte Projekt einen Beitrag im Bereich der Entwicklung von Ansätzen zur Aufbereitung großer textbasierter Datenbestände sowie der Erforschung des Sprachgebrauchs zeitgenössischer digitaler Kunstszenen, aber auch hinsichtlich Nutzung semantischer Technologien.

Die Anwendung von NLP-Verfahren für textbasierte Kommunikation in soziale Medien bedarf einiger Anpassungen an die sprachlichen Besonderheiten (Maynard 2012). Die Nutzung standardisierter Techniken ist bisher nur wenig erfolgversprechend (Gimpel 2011; Finin 2010). Bestehende Frameworks, wie das Natural Language Toolkit (NLTK, vgl. Bird et al. 2015), bieten die Möglichkeit der Implementierung eines individuellen NLP-Prozesses, bei dem verschiedene Verarbeitungsschritte modular integriert und miteinander kombiniert werden können. Für das vorliegende Projekt wurde eine Pipeline konzipiert und implementiert, die die Generierung von Annotationsebenen, begonnen mit der Tokenisierung und Part-of-Speech Tagging bis hin zur Extraktion von relevanten werkbeschreibenden Begriffen umfasst. Zur Evaluation des entwickelten Ansatzes wird ein regelbasiertes überwachtes Experiment mit einer definierten Teilmenge von 1255 Kommentaren durchgeführt. Es lässt sich feststellen, dass Emoticons und Partikeln falsch verarbeitet werden. Darüber hinaus werden auch Nomen, Verben und Adjektive, insbesondere Gerundien häufig falsch annotiert. Das Experiment zeigt, dass die konzipierte Pipeline für das vorliegende Kommentarkorpus iterativ optimiert werden muss. Der generierte Index werkbeschreibender Terminologie wird ferner für die Erweiterung einer domainspezifischen Ontologie zur Unterstützung semantischer Annotation verwendet. Hierfür wird ein Ansatz für das Lernen von Ontologien aus Texten verfolgt, wobei die ermittelten Begriffe als Kandidaten für Instanzen beschrieben werden. Als Referenzontologie wird eine auf CIDOC CRM-basierte Adaption verwendet (Hastik et al. 2013).

Dieses Projekt präsentiert einen innovativen Ansatz, um mit NLTK Kommentartexte aus Onlineforen der Demoszene zu annotieren. Das Standard-Tagset muss jedoch angepasst werden. Die Erweiterung der CIDOC CRM-basierten Ontologie auf Basis des generierten Indexes ermöglicht die semantische Beschreibung der Werke.

Bibliographie

Bird, Steven / Klein, Ewan / Loper, Edward (2015): *Natural Language Processing with Python*. NLTK Book http://www.nltk.org/book/ [letzter Zugriff 15. Februar 2016].

Farzindar, Atefeh / Inkpen, Diana (2015): *Natural Language Processing for Social Media*. San Francisco: Morgan & Claypool.

Finin, Tim / Murnane, Will / Karandikar, Anand / Keller, Nicholas / Martineau, Justin (2010): "Annotating Named Entities in Twitter Data with Crowdsourcing", in: *Proceedings of the NAACL HLT* 80–88.

Gimpel, Kevin / Schneider, Nathan / O'Connor, Brendan / Dipanjan, Das / Mills, Daniel / Eisenstein, Jacob / Heilman, Michael / Yogatama, Dani / Flanigan, Jeffrey / Smith, Noah A. (2011): "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments", in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* 42-47.

Hastik, Canan / Steinmetz, Arnd / Thull, Bernhard (2013): "Ontology based Framework for Real-Time Audiovisual Art", in: *IFLA World Library and Information Congress*. 79th IFLA General Conference and Assembly: Audiovisual and Multimedia with Cataloguing http://library.ifla.org/87/1/124-hastik-en.pdf [letzter Zugriff 15. Februar 2016].

Maynard, Diana / Bontcheva, Kalina / Rout, Dominic (2012): "Challenges in Developing Opinion Mining Tools for Social Media", in: *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at International Conference on Language Resources and Evaluation (LREC 2012)* 8.

Reyes, Antonio / Rosso, Paolo / Buscaldi, Davide (2012): "From Humor Recognition to Irony Detection: The Figurative Language of Social Media", in: *Data Knowledge Engineering*. Applications of Natural Language to Information Systems 74: 1-12.