CATMA - Eine Plattform zum kollaborativen und automatisierten Annotieren und Analysieren von Texten

,
thomas.boegel@informatik.uni-heidelberg.de
Universität Heidelberg, Deutschland

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

,
marco.petris@uni-hamburg.de
Universität Hamburg, Deutschland

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

Beschreibung

Dieser Workshop widmet sich der Textannotation und der Textanalyse mit der web-basierten Annotationsplattform CATMA (*Computer Aided Text Markup and Analysis*) (Meister et al. 2015), welche seit 2008 an der Universität Hamburg entwickelt wird. Die Bedarfe der Modellierung und Operationalisierung geisteswissenschaftlicher Konzepte und die Anwendung dieser Modelle auf Textdaten stand bei der Entwicklung von CATMA im Fokus. CATMA ist Open Source und außerdem XML / TEI-kompatibel, dadurch ist die Nachnutzbarkeit der mit CATMA erstellten Annotationen und Analyseergebnisse sichergestellt. Der Workshop wird neben einer Einführung in die Nutzung der Plattform für manuelles Annotieren auf zwei weitere – gerade bei der Umsetzung größerer Annotationsprojekte wesentliche – Aspekte genauer eingehen: Kollaboration und Automatisierung.

Im Workshop wird gezeigt werden, welche Funktionalitäten in CATMA für kollaboratives Arbeiten zur Verfügung stehen und wie das kollaborative Arbeiten unter den erschwerten Bedingungen der literaturwissenschaftlichen Praxis, wie z. B. der Polyvalenz literarischer Texte, möglich ist. Beim automatischen Erstellen von Annotationen hingegen spielen die kürzlich im Rahmen des heureCLÉA -Projektes in CATMA integrierten Möglichkeiten eine zentrale Rolle.

Ziel von heureCLÉA ist die Bereitstellung einer digitalen Heuristik zur Annotation von einfachen bis komplexen Konzepten der Narratologie. Unter einer digitalen Heuristik verstehen wir ein Werkzeug zur automatischen und semiautomatischen Annotation. Das heureCLÉA-Projekt konzentriert sich hierfür auf die Analyse temporaler Phänomene. Auf einer einfachen, an der Textoberfläche orientierten Ebene handelt es sich hierbei unter anderem um die Erkennung von Zeitausdrücken in literarischen Texten, auf einer komplexeren, metatextuellen Ebene beispielsweise um die Erkennung von Phänomenen der zeitlichen Ordnung wie Prolepse und Analepse (Gius / Jacke 2014). Dafür entwickelten wir einen auf manuellen und automatischen Annotationen basierten Ansatz, in dem die regelbasierte Extraktion und Normalisierung von Zeitausdrücken als Ausgangspunkt für Machine Learning Verfahren verwendet wurde (Bögel et al. 2015). Eine ganz wesentliche Komponente dieses Ansatzes ist das an der Universität Heidelberg entwickelte System HeidelTime (Strötgen / Gertz 2013). Sowohl die automatische Annotation von Zeitausdrücken, als auch linguistischer Oberflächenphänomene (Wortarten und Satzgrenzen), sowie Tempusannotationen sind bereits in CATMA integriert und können mit manuellen Annotationen kombiniert werden.

Im Workshop werden zunächst wesentliche Aspekte wie Modellierung, Annotation, Analyse, Kollaboration und Automatisierung mit CATMA anhand der Erfahrungen des heureCLÉA-Projektes vorgestellt. Anschließend haben die Teilnehmer_innen die Gelegenheit in einer praktischen hands-on-Session CATMA sowie HeidelTime und andere Komponenten der in CATMA integrierten NLP-Pipeline auszuprobieren. Automatische Annotationen können evaluiert werden und mit manuellen Annotationen kombiniert in die Analyse einfließen. Es kann entweder mit eigenen oder kollaborativ mit von uns zur Verfügung gestellten Texten gearbeitet werden.

Wir erhoffen uns außerdem durch den Workshop kritisches Feedback zur weiteren Verbesserung von CATMA und eine Diskussion über die Anforderungen für Textanalyse-Plattformen in verschiedenen Bereichen der Digital Humanities.

Beitragende

Alle Veranstalter_innen sind Mitglieder des heureCLÉA-Projektes. Wir haben auf zahlreichen nationalen und internationalen Tagungen und Konferenzen unsere Arbeiten zu CATMA, HeidelTime und heureCLÉA vorgestellt. Dieser Workshop baut auf Erfahrungen aus anderen Workshops zum selben Thema sowie der Einbettung von CATMA in die Lehre auf. Auch das sehr positive Feedback vergangener Workshops, z. B. zu unserem Tutorial im Rahmen der DH 2014, hat uns dazu motiviert, erneut einen Workshop anzubieten bzw. einen Antrag dafür einzureichen.

Thomas Bögel , Institut für Informatik, Universität Heidelberg,

Nach seinem Computerlinguistikstudium begann Thomas Bögel sein Promotionsstudium am Institut für Informatik an der Universität Heidelberg, wo er auch wissenschaftlicher Mitarbeiter ist. Seine Forschung beschäftigt sich vor allem mit "event extraction" und "timeline generation" sowie mit der Entwicklung von Machine Learning Systemen für die Extraktion von temporalen Relationen in narratologischen Texten.

Evelyn Gius , Institut für Germanistik, Universität Hamburg,

Evelyn Gius forscht und lehrt im Bereich der Digital Humanities mit einem Fokus auf computergestützter Textanalyse und der Hermeneutik digitaler Zugänge zu Texten. In ihrer Promotion hat sie mithilfe von CATMA an einem Korpus von Erzählungen über Arbeitssituationen untersucht, inwiefern narratologische Kategorien aus der Literaturwissenschaft für die Analyse der Konflikthaftigkeit von Alltagserzählungen genutzt werden können.

Marco Petris , Institut für Germanistik, Universität Hamburg,

Marco Petris ist Informatiker mit starker Affinität für die Geisteswissenschaften und hat von Beginn an CATMA federführend aufgebaut. Als Software Entwickler ist er in zahlreiche Projekte für die Digital Humanities involviert, wobei er sich dabei vor allem um die Konzeption und Implementierung kümmert.

Jannik Strötgen , Max-Planck-Institut für Informatik, Saarbrücken,

Bevor Jannik Strötgen als Postdoc zum MPI wechselte, studierte er in Heidelberg Computerlinguistik und promovierte und arbeitete am Institut für Informatik der Universität Heidelberg. Im Rahmen seiner Dissertation beschäftigte er sich vor allem mit Informationsextraktion sowie Information Retrieval und begann die Entwicklung von HeidelTime, einem frei verfügbaren Temporal Tagger, der für verschiedene Domänen und Sprachen geeignet ist.

Kapazität und Ausstattung

Die Teilnehmerzahl ist auf 20 Personen begrenzt. Jede_r Teilnehmer_in braucht einen Laptop (ein Tablet PC reicht nicht aus!) und ein Google Mail Konto für den CATMA Login.

Fußnoten

1. heureCLEA ist ein BMBF-gefördertes ehumanities Projekt zwischen der Universität Hamburg und der Universität Heidelberg (Gertz / Meister 2016).

Bibliographie

Bögel,Thomas / Strötgen, Jannik / Gertz, Michael (2015): „A Hybrid Approach to Extract Temporal Signals from Narratives“, accepted at: *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL’15)*, Duisburg-Essen, Germany.

Gertz, Michael / Meister, Jan Christioph (2016): *heureCLÉA*. Collaborative Literature exploration & annotation. Hamburg: Universität Hamburg http://heureclea.de/ [letzter Zugriff 08. Oktober 2015].

Gius, Evelyn / Jacke, Janina (2014): „Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets“. Hamburg 2014 http://heureclea.de/publications/guidelines.pdf [letzter Zugriff 08. Oktober 2015].

Meister, Jan Christoph / Gius, Evelyn / Petris, Marco / Meister, Malte / Jacke, Janina (2015): *CATMA*. Computer Aided Textual Markup Computer Aided Textual Markup & AnalysisAnalysis. Hamburg: Universität Hamburg http://www.catma.de/ [letzter Zugriff 08. Oktober 2015].

Strötgen, Jannik / Gertz, Michael (2013): „Multilingual and cross-domain temporal tagging“ in: *Language Resources and Evaluation* 47, 2: 269-298.