ATHEN – Ein Werkzeug zur Annotation von Textkorpora

Linguistische Ressourcen stellen im Zeitalter der datengetriebenen Algorithmen den wichtigsten Bestandteil dar, es gilt das Motto „je mehr Daten desto besser“. Dies legt nahe, einen möglichst effizienten Weg zu finden, diese literarischen Ressourcen zu erstellen.

Standardmäßig beginnt ein linguistisches Projekt etwa mit Hypothesen oder Aufträgen, die ein gewisses Maß an Metainformationen in textuellen Ressourcen erfordern. Aus den Metainformationen wird ein abstraktes Annotationsschema entwickelt mit dessen Hilfe letztlich, zunächst manuell, Annotationen erstellt werden sollen.

ATHEN (Annotation and Text Highlighting Environment) ist in der Lage, aufgrund des sehr mächtigen Apache UIMA Frameworks, beliebige Annotationsschemata, die in ein Apache UIMA Typesystem abgebildet werden können, zu importieren und die manuelle Annotation bestmöglich zu unterstützen. ATHEN setzt dabei auf ein gutes visuelles Feedback. Der Nutzer kann – je nach aktuellem Annotationsprojekt – zwischen einer Reihe verschiedener Darstellungsmöglichkeiten wählen.

Darüber hinaus unterstützt ATHEN in besonderer Weise das Annotieren folgender Vorhaben:

- Named Entities
- Koreferenzen
- Relationen
- Direkte Reden inklusive Sprecher und Angesprochener Entität
- Szenen
- Konstituenzgrammatiken
- Dependenzgrammatiken

ATHEN besitzt neben dem Annotieren von textuellen Dokumenten noch weitere Features:

- OWL-Support zum Erstellen und Annotieren von Ontologien
- Apache Lucene-Support, um effizient Suchanfragen über zuvor erstellte Annotationen (egal ob manuell oder automatisch) zu beantworten.
- Die Erstellung von Figurennetzwerken
- Das Annotieren von Bilddaten durch spezielle Annotationen in Apache UIMA
- Die Möglichkeit Annotationen (etwa mehrerer Annotatoren) gegenüber zu stellen, mit automatischer Anzeige der Übereinstimmung.
- Eine vollständige Unterstützung des Prozessmodells medizinischer, ontologiebasierter Informationsextraktion.
- Konfigurierbare, automatische Vorverarbeitung und Anwendung UIMA basierter Analysis Engines

- Das Konvertieren zwischen verschiedenen Formaten
- Eine Erweiterbarkeit, ohne den eigentlichen Code verändern zu müssen durch sein flexibles OSGI-Design
- Eine selbst konfigurierbare Oberfläche durch das Einstellen der sichtbaren Views.

Der Vortrag wird auf einige der genannten Features eingehen und anhand von Videos zeigen, wie ein bestimmtes Annotationsvorhaben mit ATHEN umgesetzt werden kann.

Markus Krug

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik

Julius-Maximilians-Universität Würzburg