Annotation and beyond – Using ATHEN Annotation and Text Highlighting Environment

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Deutschland

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Institut für Deutsche Sprache, Deutschland

Weimer, Lukas
lukas.weimer@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Universität Würzburg, Deutschland

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
puppe@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Goals of the workshop

The workshop presents ATHEN [1] (**A**nnotation and **T**ext **H**ighlighting **En**vironment), an extensible desktop-based annotation environment which supports more than just regular annotation. Besides being a general purpose annotation environment, ATHEN supports indexing and querying support of your data as well as the ability to automatically preprocess your data with Meta information. It is especially suited for those who want to extend existing general purpose annotation tools by implementing their own custom features, which cannot be fulfilled by other available annotation environments. On the according gitlab, we provide online tutorials, which demonstrate the use of specific features of ATHEN.

Related Work

We compare ATHEN to three web-based and four desktop applications in 12 categories by adapting most criteria defined by Neves and Leser (Neves / Leser 2012) to compare different annotation tools:

1. Availability and up-to-dateness of the documentation
2. Active development at the present time
3. Source code for download
4. Complexity of system requirements
5. Interoperability by supporting certain formats
6. Support of different annotation layers
7. Support of NLP-preprocessing to speed up manual annotation
8. Support of visualization
9. Support of self-learning systems to speed up manual annotation
10. Support of querying annotated data
11. Possibility to do an inter-annotator-agreement, this is important for projects, in which more than one annotator labels the same documents
12. Extensibility

We explicitly do not want to compare subjective features like usability or how the annotations are presented.

Annotator / Criterion	ATHEN	CATMA	Web-Anno	BRAT	UAM	MMAX2	Know-tator	WordFreak
Documentation	✓	✓	✓	✓	✓	✓	✓	Homepage
Active development	✓	✓	✓	✓	✓	×	×	×
Open Source	✓	✓	✓	✓	×	✓	✓	✓
System requirements	Java	Webbrowser	Java, Web-browser	Python	Win, Mac	Java	Java, Protégé	Java
Supported formats	.txt, .xmi, .xml (incl. page xml, .tei)	.tei, .txt	.CoNLL, .tcf, .teicph, .xmi	.ann, .txt,	.txt, .xml	.txt, .xml	.txt, .xml	.ace, .ptb, .muc, .txt
Supported annotation layer	Customizable	Customizable	Customizable	Entities, Relations, Events	Customizable	Markables, Attributes and Relations in between	Ontology-based	Depending on plugin
Preprocessing	UIMA Analysis Engines	using heuréCLEA	×	Sentences, tokens	POS-Tagging, Syntactic parsing	Tokenization	×	Depending on plugin
Visualization	Social networks	✓	×	×	×	×	×	×
Automation	RUTA annotation support	Machine Learning	Machine Learning	×	Query Based	×	×	×
Query language for corpus analysis	Apache Lucene Support	Query	×	×	Search and statistics	Query	×	×
Inter-annotator agreement	Interactive side by side	×	Curation View	Side by side	×	×	✓	×
Extensibility	Runtime/ Code	Code	Code + Tutorial	Code	×	Code	Code	Code/ Plugin

Table 1: Comparison of three web-based and four desktop applications with ATHEN in twelve categories. No tool excels in every category.

All of the listed tools have an accessible documentation, either web-based or as a PDF, available for download. Besides UAM (O'Donnell 2008), every other application is listed as open source, so at least extensions based on code level can be made. WebAnno (Yimam et al. 2013) is the only application having a tutorial supporting a new developer to make changes in their project. ATHEN stands out in the sense that extensions to its UI can be made at runtime, therefore easing the process of adding functionality to it. WebAnno supports the largest number of formats and comes with a machine learning based automatic annotation, however lacks integrated NLP-preprocessing. CATMA (Meister 2017) is the only project that has a very good visualization component and also supports TEI-XML (Wittern et al 2009), the

unspoken standard of text processing. Being a standalone web application, CATMA itself does not support NLP-preprocessing. ATHEN comes with the support of the execution of UIMA analysis engines, accessible from web repositories or a local repository, giving the user a chance to integrate her custom-made annotators. Four tools, ATHEN, UAM, MMAX2 (Müller / Strube 2006) and CATMA feature an integrated query language which helps to analyze existing corpora. Most tools allow the annotation of user-defined annotation schemas earning therefore the title “generic annotation tool”. Alongside UAM, ATHEN supports the annotation based on queries, while UAM defines its own language, ATHEN supports the annotation using Apache UIMA Ruta (Kluegl et al. 2016) rules. Three of the listed tools, MMAX2, Knowtator (Ogren 2006) and WordFreak (Morton / LaCivita 2003) are currently no longer in active development.

Brief technological description of ATHEN

ATHEN is a Java-based desktop application with the vision to be extensible. Therefore, it makes use of the flexible plugin architecture of the eclipse **R**ich- **C**lient-**P**latform (RCP) [2]. Internally, it is built around Apache UIMA, which means incoming data is automatically converted into the UIMA specific Common Analysis Subject (CAS) architecture. Working with UIMA allows the integration of standalone analysis engines, which can be used to preprocess data and speed up manual annotation. The use of Apache Lucene enables ATHEN to create an index comprising documents, as well as their annotations, which results in queries that can answer questions based on text and meta information in real time. With the ability to execute Apache UIMA Ruta one can even create queries of far higher complexity. On top of that, ATHEN features OWL-Support, which allows the definition of an ontological annotation schema in a machine- readable format. Using Apache UIMA internally allows ATHEN to even address more complicated input. Currently ATHEN supports the annotation of image regions, based on user defined polygons.

Program of the workshop

The program is split into four sections:

1. Introduction to ATHEN and distinguishing from other existing annotation environments.
2. Working with ATHEN, which contains the definition of scenarios and annotating sample documents.
3. Utility of ATHEN (beyond regular annotation), which addresses the following topics:
    1. Defining an annotation schema using OWL
    2. Preprocessing texts based on Apache UIMA Analysis Engines
    3. Creating and executing queries based on Apache Lucene
    4. Annotating images with ATHEN
4. Extending ATHENs functionalities and adapting it to your needs (developer specific)

The first section is a presentation which shows the main differences between the existing annotation tools. The second section defines an ordinary annotation scenario and it is used to introduce the participants to the general-purpose annotation view of ATHEN. Afterwards, for tasks to which ATHEN has special support (annotating character references and their coreferences, annotating direct speech and their speaker) an introduction to the special purpose views of ATHEN is given.

The third panel introduces the participants to the functionality of ATHEN beyond regular text annotation. It starts with the definition of an OWL ontology (and its utilization for texts). This is centered on relation detection of character references, as well as an attribution of those references.

To speed up manual annotation it is helpful to have it preprocessed with existing tools. The task definition is then changed from pure annotation to an application with a consecutive correction of the output of the automatic engines. In this context, Nappi, a submodule of ATHEN is presented and it is shown how to define, execute and integrate custom analysis engines.

The next part is dedicated towards extracting knowledge from annotated data, for this purpose, an Apache Lucene Index is created using ATHEN and is queried in a live fashion. This feature allows rapid insight into an existing corpus and enables the user to answer their own hypothesis.

The tutorial continues with the presentation of how images can be annotated with polygon-based annotations to show, that ATHEN is not only limited to textual resources.

The last part is directed towards Java developers who are interested in developing their own annotation component.

Each section starts with a set of slides which introduce the features in focus and presents the participants with one or more tasks that can be fulfilled by using ATHEN.

Requirements

The participants need their own laptops with an active internet connection. The number of participants is limited to 15 to 20. The last section requires knowledge of Java. Data which is necessary for the tutorials will be hosted on our own server and will be made accessible for download.

Research projects

ATHEN is mainly developed in the context of the project Kallimachos at the University of Wuerzburg. Its main purpose was to support the annotation process of DROC (**D**eutscher **RO**man **C**orpus). Currently automatic creation of literary interaction networks, automatic genre detection of novels and sentiment analysis in literary novels are in the focus of interest.

An extension to ATHEN was made in the project "Redewiedergabe" to manually annotate different forms of speech, thought and writing representation (STWR). These annotations will then be used to train an automatic recognizer for STWR.

Fußnoten

1. https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen
2. A web version of ATHEN with its major features is available at: https://webathen.informatik.uni-wuerzburg.de

Bibliographie

Kluegl, Peter / Toepfer, Martin / Beck, Philip-Daniel / Fette, Georg / Puppe, Frank (2016): "UIMA Ruta: Rapid Development of Rule-based Information Extraction Applications", in: *Natural Language Engineering 22.1* 1-41.

Meister, Jan Christoph / Gius, Evelyn / Jacke, Janina / Petris, Marco: *CATMA 5.0*. http://catma.de/ [Accessed September 22, 2017].

Morton, Thomas / LaCivita, Jeremy (2003): "WordFreak: an open tool for linguistic annotation", in: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4* 17-18.

Müller, Christoph / Strube, Michael (2006): "Multi-level annotation of linguistic data with MMAX", in *: Corpus technology and language pedagogy: New resources, new tools, new methods 3* 197-214.

Neves, Mariana / Leser, Ulf (2012): "A survey on annotation tools for the biomedical literature", in: *Briefings in Bioinformatics 15.2* 327-340.

O'Donnell, Mick (2008): "Demonstration of the UAM CorpusTool for text and image annotation", in: *Proceedings of the 46th annual meeting of the Association for computational linguistics on human language technologies: Demo session* 13-16.

Ogren, Philip V. (2006): "Knowtator: a Protégé plug-in for annotated corpus construction", in: *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations* 273-275.

Stenetorp, Pontus / Pyysalo, Sampo / Topi#, Goran / Ohta, Tomoko / Ananiadou, Sophia / Tsujii, Jun'ichi (2012): "BRAT: a Web-based Tool for NLP-Assisted Text Annotation", in : *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* 102-107.

Wittern, Christian / Ciula, Arianna / Tuohy, Conal (2009): "The making of TEI P5", in: *Literary and Linguistic Computing 24.3* 281-296.

Yimam, Seid Muhie / Gurevych, Iryna / de Castilho, Richard Eckart / Biemann, Chris (2013): "WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations", in: *Proceedings of ACL-2013, demo session, Sofia, Bulgaria* 1-6 .