WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations

Seid Muhie Yimam
Chris Biemann
Richard Eckart de Castilho
Benjamin Milde
Darina Benikova

November 05, 2013
Overview

• Introduction
• Architecture and workflow
• System requirements and installation
• Project definition
• Annotation/Correction/Curation
• Project monitoring
• Crowdsourcing
• Live demo
• Feedback
Introduction

• Linguistic annotation is an addition of meta data to parts of a document (word, phrase, sentence,...)
• Corpora annotated with linguistic information are required to develop robust statistical NLP systems.
• Building such corpora, however, is an expensive, labour-intensive, and time-consuming work.
• An annotation tool that will facilitate development of such corpora is needed.
• Previous annotation tools
  – mostly specific to one annotation layer
  – mostly platform-dependent
  – mostly single-user, no administrative support
WebAnno: Overcoming previous limitations

Properties of the WebAnno annotation tool:

- **Flexibility**: several annotation layers, several formats, extensible
- **Web-based**: available everywhere, no installation effort
- **Open source**: ensuring maintainability and usability
- **Quality and User Management**: Curation and Project setup interfaces
- **Crowdsourcing**: scaling out to large distributed workforce
- **Pre-annotated and un-annotated documents**: new annotations and correcting automatic annotations
Technologies WebAnno built on

- Java EE
- Hibernate, Spring, Apache Wicket, Apache UIMA
- MySQL
- The brat front end
Project architecture

- Admin
- Curator
- Annotators
- Crowd

Server:
- Use case correct and retrain
- Task definition
- BRAT + Choice Adjudication
- BRAT full text annotation
- Crowd-Server: Task generation, control, result aggregation

Data Layer:
- UIMA CAS Persistent Storage
- Exchange Formats with WebLicht and others

Users:
- Admin
- Task definition
- BRAT + Choice Adjudication
- BRAT full text annotation
- Crowd Frontend single annotation

Annotators Crowd Frontend
Workflow of WebAnno projects

Project setup
- Import documents
- Define guidelines
- Define types/tagsets

Annotation of documents
- Assign workload
- Monitor progress
- Evaluate quality

Monitoring
- Compare annotations
- Merge annotations
- Correct annotations

Curation

Annotators
Crowd

Manager
Curator

EXPORT FINAL DATASET
System requirement

- Application Server (Tested with Apache Tomcat)
- Database – MySQL
- Browser - compatible with chrome and safari
Installation

• Directory to store annotation/source documents
  – /srv/webanno
  – Property files such as data base settings and file formats
    • settings.properties
    • formats.properties

• Tomcat
  – JAVA_OPTS should have the pointer to webanno home
    • JAVA_OPTS="-Dwebanno.home=/srv/webanno"

• MySQL
  – Just create the database
    – >create database webanno
Installation

• Installing from Source
  – Check out the code
    – `svn checkout https://webanno.googlecode.com/svn/trunk/ webanno`
    – Unzip and go to the webanno directory
    – `mvn clean install`
    – Get the war file, `webanno.war` and put in webapps of tomcat

• Getting the latest war file
  – Usually from google code Downloads page,
  – Get the version you want and put it in webapps of tomcat
First time login

• No user is found, hence automatically create an admin user (username as admin and password as admin)
  – Should be changed for security reason!

07.11.2013
Welcome page

• Super admin
  – Can see all of the pages

• Admin
  • Can see Curation, projects, and crowdSource pages

• User
  – Can see only the annotation link if user is assigned to any project

• Curator
  – Can see curation, correction, and monitoring links
User management

• Add a user
• Set role
• Delete user (disabled), what happened to the existing annotations.
# User management

## Global Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Privileges</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROLE_ADMIN</td>
<td>Create projects, users</td>
</tr>
<tr>
<td>ROLE_USER</td>
<td>Annotate, curate, monitor, manage project, crowd source</td>
</tr>
<tr>
<td>ROLE_REMOTE</td>
<td>Create project into WebAnno through web services</td>
</tr>
</tbody>
</table>

## Project Roles

<table>
<thead>
<tr>
<th>Roles</th>
<th>Privileges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin</td>
<td>Manage project, Monitoring, crowdsourcing</td>
</tr>
<tr>
<td>User</td>
<td>Annotate a document</td>
</tr>
<tr>
<td>Curation</td>
<td>Curate annotations, monitor project</td>
</tr>
</tbody>
</table>
Project definition

- Project description
- Adding users
- Adding documents
- Defining tagsets and tags
- Adding project guidelines
Project definition

- name & desc.
- user roles
- add/edit tagsets
- import docs
- upload guidelines
- Export (to file system, CLARIN-D)
Project definition

• Define project type
  – Annotation project
  – Correction project

• Exporting
  – The whole project
  – Only curated documents
  – Send to the CLARIN infrastructure

• Import
  – The whole Project
    • Annotated document
    • Curated document
    • Tags and document meta-data from the database
Adding a user

• Add new user from user lists
• Modify the default role (user) accordingly
Adding documents

• Select a file or multiple files of the same type (TCF, XMI, TSV, Text) and import
Adding tagsets

• Current implementation, predefined types
  – POS
  – Dependency
  – Lemma
  – Named Entity
  – Coreference

• Export and import tagsets
  – JSON format
  – TSV
Tagset editor

- **Tagsets**
- **Desc. Of tagset**
- **Export tagset as JSON or TAB-sep**
- **Import tagset**
- **Tags**
- **Desc. of tag**
Annotation

- project navigation
- navigation within document
- settings, configures display
Annotation

• Span Annotation
  – Single Token
  – Multiple Token
  – Do not cross a sentence
  – Named Entity, POS, Lemma, Coreference Type

• Arc Annotation
  – Between span annotation
  – Reversible arcs
  – Dependency parsing

• Chain annotation
  – Between spans (mentions)
  – Uni-directional arcs
  – Coreference chain
Annotation

• Pagination of Large document
  – Based on window size (number of sentences per window)
  – Navigate to first, last, previous and next pages
  – Go to specific sentence number
  – Scroll while annotating
Admin can open all documents

All projects

All documents in a project
But anno1 can open only her own Projects for anno1 as a user role

Documents anno1 allowed to work in the selected project
Named Entity and Co-reference

Named Entity

1. AC Milan chief executive Adriano Galliani believes that Mario Balotelli is destined for greatness.

Coreference

2. Balotelli arrived back in Italy at the end of the January transfer window, but after starting his career with rivals Inter Milan, the 22-year-old has swapped the blue half of Milan for the red, and signed a long-term deal with the AC Milan.
Language support
Exporting annotated document

File formats
Correction

• Get automatically annotated document
  – Clear all existing annotation
  – Upper half of the window for free annotation; lower half of the window to display the auto-annotated document
Correction

User’s new annotation

Copied annotation

Auto-generated annotations
Curation

- Curate annotation conflicts among users
- For every annotation type, detect discrepancy in annotation
Curation

highlight sentences with disagreement

display annotators, color-coded agreement
Monitoring

- Percentage (task accomplishment)
- Number of documents finished
- Kappa
  - Two paired kappa (Per user, document and type)
Monitoring

anno1 finished 2 documents
anno1 has finished annotation!
Crowd-sourcing

- Crowdsourcing German named entity recognition with WebAnno and Crowdflower.com
2-stage named entity recognition

• Implemented on crowdflower.com as two separate tasks:
  • First task: workers are asked to identify possible spans that could be any named entity
  • In a second task, workers identify the type of such a span
Recognizing spans:

- Workers select words of interest with their mouse:
First results:

Sentence level agreement (between workers): 75%

- Price recommendation: 13 dollar cents per 5 judgments
- 500 sentences
  - 5 judgments per sen. = 325 $
  - 3 judgments per sen. = 195 $
  - Total time: 34 hours
Selecting type of named entity:

Workers select type with radio buttons:

Im folgenden Satz: Auch *von Estorff* betont, dass im Fall Haesefeld keine Entscheidung übers Knie gebrochen werden dürfe: "Die Formalitäten wie die Einwände arbeitet der alte Rat ab, den Rest macht der neue."

Ist "*von Estorff*" eine: *(required)*
- Person
- Organisation
- Ort
- Etwas anderes
- Gewöhnliches Wort
- Falsch markiert / nicht die ganze Einheit markiert:
- Nicht sicher:

Kommentar:

Optional
First results:

Sentence level agreement (between workers): 85%

- Price recommendation: 5 dollar cents per 5 judgments
- 500 entities:
  - 5 judgments per sen. = 125 $
  - 3 judgments per sen. = 75 $
    - Total time: 10 hours
WebAnno integration

• Both jobs can easily be created and populated with data on the crowdsourcing page directly in webanno
• Source documents will be crowdsourced
• Curated documents can be selected as gold data
• Named entities will be aggregated and displayed as crowd user
• Payment and start of job via crowdflower.com platform
WebAnno in progress

- Free/unrestricted annotation layers
- Automatic annotation
- File type validation while importing
- Multiple file format import/export
- Replace the annotation dialog with menu structure
WebAnno in Google Code

• Including
  – Installation instruction
    • Quick installation
    • Developer installation
  – User’s Guide
  – Upgrade procedure
  – FAQ
  – Troubleshooting
  – Sample data
• Available [here](#)
Live demo

• Go through all steps together
  – Create user for attendee
  – Add users to existing project
  – Add tagsets if needed
  – Annotate documents
  – Curation
  – monitoring