Interconnecting the LAPPS Grid and PubAnnotation for Rapid Development of Corpus Annotation with Machine Learning in the Loop

Nancy Ide, Keith Suderman
Department of Computer Science, Vassar College
Poughkeepsie, New York USA

Jin-Dong Kim
Database Center for Life Science (DBCLS) Research Organization of Information and Systems
Wakashiba, Kashiwa-shi, Chiba, Japan

LAPPS Grid

https://www.lappsgrid.org

- Provides seamless access to a wide range of NLP tools and resources
  - popular public NLP tools, tools for biomedical and clinical data analysis, many others
  - hundreds of additional multilingual tools and data sources through federation with Kyoto University's Language Grid and EU-CLARIN frameworks (WebLicht and LINDAT/CLARIN)
- All tools and resources are *interoperable*
- Galaxy front end provides a "plug-and-play" workflow environment
- Run from the web, on a user's laptop or desktop, in the cloud, or as a self-contained docker image

PubAnnotation

http://www.pubannotation.org/

- A repository of text annotations on literature of life sciences
- Annotations registered in PubAnnotation are aligned to the canonical text from PubMed or PMC
  - all annotations thus linked to each other through canonical texts
  - annotations accessible and searchable through standard web protocol
- Includes TextAE, a powerful and easy-to-use Javascript app for text annotation and visualization

Collaboration

Last year, LAPPS Grid developers and developers of PubAnnotation began integration of services and resources provided by each

- LAPPS Grid users can
  - register and thus share annotations in the PubAnnotation registry
  - access publications available through PubAnnotation (PubMed, PMC)
  - invoke TextAE from within the LAPPS Grid
- PubAnnotation users can:
  - access and apply LAPPS Grid tools from within the PubAnnotation environment
- Automatically convert between PubAnnotation's JSON format and the LAPPS Grid JSON-LD format so that interoperability is seamless from the point of view of the user

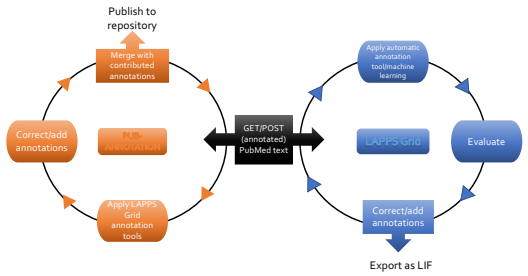
Result

- Greatly enhanced ability to annotate scientific publications and share results
  - users can easily apply automatic annotation tools and manually correct annotations in an iterative "human-in-the- loop" process of refinement
  - useful for the creation of training data for machine learning (esp. semi-supervised approaches such as active learning)

Overall Goal

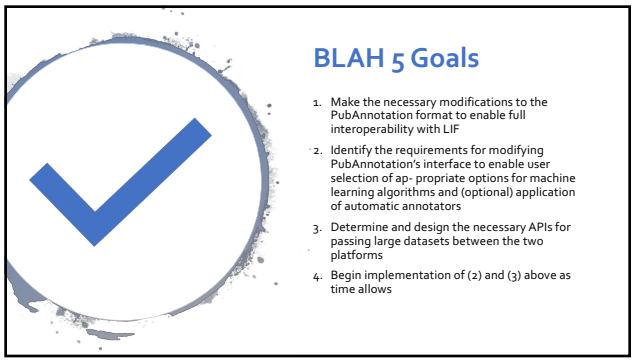
- Continue to implement the interoperability layer between LAPPS Grid and PubAnnotation to provide an eco-system consisting of manually annotated data (in PubAnnotation) and machine learning algorithms (in LAPPS Grid)
- Enable an iterative cycle of annotation and learning e.g.:
  - An annotation project on PubAnnotation includes documents, some of which are already annotated
  - The maintainer of the project chooses a machine learning algorithm from among a range of options available on the LAPPS Grid
  - The annotated documents in the project are sent to the LAPPS Grid and used to train the language model
  - The model is used to annotate the remaining documents in the project
  - The annotator may manually correct some of the automatic annotations, retrain the model, then automatically re-annotate the documents using software available in the LAPPS Grid, as many times as necessary until the desired results are achieved

Envisioned Interaction of PubAnnotation and the LAPPS Grid



Remaining Tasks

- Implement means to easily identify and pass large bodies of text between the two platforms
- Design and implement a user interface to provide an efficient and effective user platform



BLAH 5 Goals

1. Make the necessary modifications to the PubAnnotation format to enable full interoperability with LIF
2. Identify the requirements for modifying PubAnnotation's interface to enable user selection of ap- propriate options for machine learning algorithms and (optional) application of automatic annotators
3. Determine and design the necessary APIs for passing large datasets between the two platforms
4. Begin implementation of (2) and (3) above as time allows