A Multi-Platform Annotation Ecosystem for Domain Adaptation

Richard Eckart de Castilho♣, Nancy Ide♠, Jin-Dong Kim♡,
Jan-Christoph Klie♣, Keith Suderman♠
♣Ubiquitous Knowledge Processing (UKP) Lab, Department of Computer Science,
Technische Universität Darmstadt, Darmstadt, Germany
♠Department of Computer Science, Vassar College, Poughkeepsie, New York USA
♡Database Center for Life Science, Research Organization of Information and Systems,
Kashiwa-shi, Chiba, Japan

Abstract

This paper describes an ecosystem consisting of three independent text annotation platforms. To demonstrate their ability to work in concert, we illustrate how to use them to address an interactive domain adaptation task in biomedical entity recognition. The platforms and the approach are in general domain-independent and can be readily applied to other areas of science.

1 Introduction

The rapidly growing appearance rate of biomedical publications has increased interest in applying natural language processing (NLP) and machine learning (ML) technologies to navigate the massive volumes of biomedical literature. In particular, the use of *text annotation* to better automate knowledge extraction and identify relevant information in the literature has become an increasingly major activity over the past decade.

Numerous platforms and frameworks that support text annotation have been developed, including the General Architecture for Text Engineering (GATE (Cunningham et al., 2013)), CLARIN WebLicht (Hinrichs et al., 2010), the Language Applications (LAPPS) Grid (Ide et al., 2014), OpenMinTeD (Labropoulou et al., 2018), and several systems based on the Unstructured Information Management Architecture (UIMA (Ferrucci et al., 2009)), e.g. ARGO (Rak et al., 2013), Apache cTAKES (Savova et al., 2010), DKPro Core (Eckart de Castilho and Gurevych, 2014). However, due to factors such as the often highly domain-specific vocabularies in specialized areas of science, these frameworks are rarely usable out-of-the-box. As a result, scholars interested in mining publications may spend considerable effort to adapt existing annotation tools and resources to their particular domains of research (e.g., tune them to domain-specific terminology), a process referred to as *domain adaptation*.

Machine-assisted interactive annotation (also known as *human-in-the-loop* annotation) is a recognized means to support domain adaptation, by enabling the rapid creation of benchmark annotation data for specialized domains, which can be used for training or adapting annotation models and evaluating their performance. This process requires several capabilities, including ready access to (1) relevant document repositories, (2) retrainable NLP tools (e.g., named entity recognizers), and (3) sophisticated annotation editors that integrate retraining into the interactive annotation process. However, because all of these capabilities are not available within any single text mining platform, the researcher must use multiple platforms and tools. And although tools and resources may be interoperable within a single platform, combining tools and resources across platforms can demand substantial computational expertise.

One approach to solve this problem would be to develop a monolithic framework that incorporates all of the requisite functionalities. Our solution is instead to interconnect three independently developed platforms, each of which supports some aspect(s) of the domain adaptation process, but none of which provides the entire suite of required tools and resources. This necessitates adaptations to achieve *interoperability* among them–i.e., to be able to exchange data among the platforms without the need for explicit conversion.

In this paper, we describe three platforms that constitute our annotation ecosystem, as background for a demonstration of their ability to work in concert to provide easily usable means to adapt NLP processes to specific domains. Our focus is on the use of the ecosystem to address text mining in the biomedical domain, but the strategies outlined are readily applied to other areas of science.

Proceedings of the 13th Linguistic Annotation Workshop, pages 189–194
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

2 Platforms

This section briefly introduces the three platforms comprising our ecosystem (Figure 1) . Each represents a particular class of systems: a repository for annotated corpora, an NLP services platform, and an interactive annotation platform. These are introduced as platforms and not as tools as they are designed as open and extensible software systems. All are open source software and users can set up their own installations, e.g. for their own project, lab, or community. Some also run a *canonical* instance accessible to any registered user.

***PubAnnotation* (Kim and Wang, 2012)** takes on the role of the annotation repository in our ecosystem. It links all contributed annotations through references to canonical texts. It also supports annotation development coupled with *PubDictionaries*, a similarly open repository of dictionaries (term lexicons, etc.) to which users can add by registering their own dictionaries or modifying those already in the repository; as well as *TextAE*, a browser-based visualizer/editor for text annotation. The service-oriented architecture makes it easy for end-users to customize annotation tools by engaging in the annotation process from start to finish. It consists of a collection of web services and web clients that can interact with other systems through REST APIs and a JSON-based data format. The SPARQL standard is supported and allows searching the linked annotations.

The *LAPPS Grid* (Ide et al., 2014) acts as the NLP services platform in our ecosystem. It provides a large collection of NLP tools exposed as web services, together with a variety of commonly used resources (e.g., gold standard corpora). The services and resources are made available via a web-based workflow development engine[1], directly via SOAP calls, and programmatically through Java and Python interfaces. All tools and resources in the *LAPPS Grid* are rendered mutually interoperable via transduction to the JSON-LD LAPPS Grid Interchange Format (*LIF* (Verhagen et al., 2016)) and the Web Service Exchange Vocabulary (*WSEV* (Ide et al., 2016)), both designed to capture fundamental properties of existing annotation models in order to serve as a common *pivot* among them.

***INCEpTION* (Klie et al., 2018)** contributes interactive annotation functionality to the ecosystem. The platform can be configured for different annotation tasks through a configurable annotation schema supporting span and relation annotation that can carry different kinds of attributes (string, numeric, boolean, etc.). It connects to external document repositories in order to search and import documents for later annotation. Automatic *recommenders* provide annotation suggestions by connecting to external NLP services or by using internal machine learning libraries. To support domain adaptation, the suggestions can be improved as the user interactively reviews and corrects them. Domain-specific vocabularies can be accessed from external SPARQL endpoints or be managed in an internal RDF knowledge base. By supporting common formats and standards for annotation representation and knowledge representation, *INCEpTION* offers a high level of interoperability. Through its remote API, it can be integrated into external workflows. The implementation is internally using the UIMA CAS (Götz and Suhre, 2004) data model.

To create a domain adaptation ecosystem from these three independent platforms, it is necessary to establish *cross-platform interoperability*, i.e., the ability to exchange data consisting of text and associated annotations among them. This means that the data must be mutually *understandable* at the *data level* (model and schema), either directly or via trivial conversion. It must also be possible to appropriately utilize data from the other platforms within the constraints of their respective architectures. In the present paper, we focus on the cross-platform scenario and on the possible actions that can be taken, while a detailed description of the challenges for interoperability among the three platforms at a more technical level and the implemented solutions is provided by Eckart de Castilho et al. (2019).

3 Domain Adaptation for Biomedical Publications

A principal requirement for effective information mining from biomedical texts is the identification of biologically and clinically relevant concepts, e.g., genes and gene products, diseases, and treatments, in the vast body of available data. Domain adaptation for biomedical texts therefore centers around the development and refining of applications for *named entity recognition* (NER), for which numerous freely available tools exist.

[1] http://galaxy.lappsgrid.org

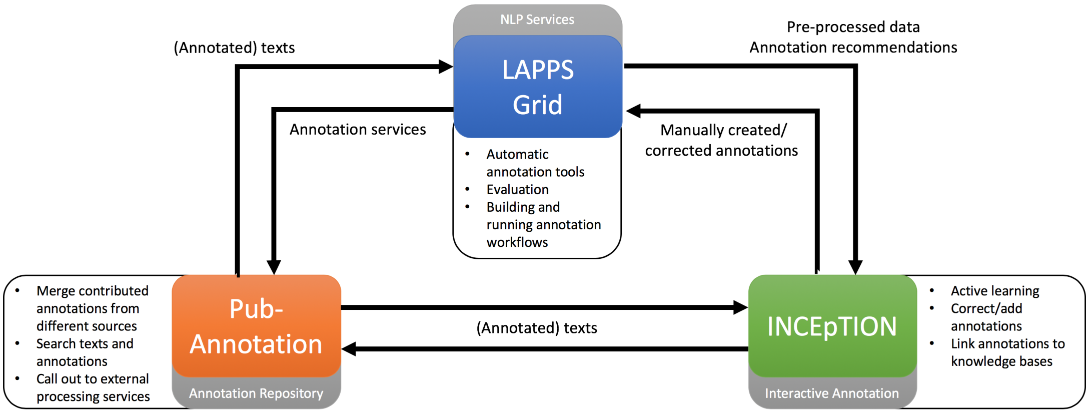
NLP Services
LAPPS Grid
Automatic annotation tools
Evaluation
Building and running annotation workflows
(Annotated) texts
Annotation services
Pre-processed data
Annotation recommendations
Manually created/ corrected annotations
Merge contributed annotations from different sources
Search texts and annotations
Call out to external processing services
Pub-Annotation
Annotation Repository
(Annotated) texts
INCEpTION
Interactive Annotation
Active learning
Correct/add annotations
Link annotations to knowledge bases

Figure 1: High-level interactions in the tripartite annotation ecosystem

	LAPPS Grid	***PubAnnotation***	***INCEpTION***
BUILD	Build a collection of texts using the *PubAnnotation* datasource or another *LAPPS Grid* datasource	Build a collection of texts from PubMed or PMC	Search an external repository (e.g. *PubAnnotation*) and selectively import relevant texts
ANNOTATE	Perform automatic annotation using one of the *LAPPS Grid*'s NER services	Call out to one of the registered annotation services, e.g. *PubDictionaries* or *LAPPS Grid*'s NER services	Import documents pre-annotated e.g. by *LAPPS Grid* services or *PubAnnotation*
EVALUATE	Compare to another corpus using the *LAPPS Grid* Open Advancement evaluation tools	Compare to another corpus registered in *PubAnnotation*-or- search results using the SPARQL- or keyword-based search interfaces	Compare annotations between users, compute inter-annotator agreement, curate results
REVISE	Edit the annotations using *TextAE* -or- edit dictionary entries externally and re-import to the *LAPPS Grid* for input to dictionary-based NER	Modify/add to the dictionary using *PubDictionaries*-or- edit the annotations using *TextAE*	Edit annotations in *INCEpTION*, optionally assisted by automatic annotation suggestions generated by an embedded ML- or dictionary-based approach, or by calling an external service (e.g., a *LAPPS Grid* NER service)
RE-TRAIN	Re-execute a machine learning algorithm with the newly annotated data, either within the Galaxy history or via direct call to the web service	n/a	When the user makes an edit, automatically retrain embedded approaches or external services if they support it.
REPEAT	Re-execute the appropriate process in the Galaxy history or via direct call to the web service	Re-execute the *PubDictionaries* NER service with revised dictionaries	Re-training and re-processing happens automatically, coupled with updated performance indicators (e.g., F-score)

Table 1: Comparison of supported activities within and across the platforms

Even given the several NER tools and frameworks that have been developed with biomedical entities in mind, including for example the Genia tagger (Tsuruoka et al., 2005), GOST tagger (El-Haj et al., 2018), Termine,[2] the Penn BioTagger[3] (Jin et al., 2006), and OGER++ (Furrer et al., 2019), results are rarely comprehensive and reliable enough to be immediately usable for serious text mining. More importantly, such tools typically cover only very general categories of bioentities, often miss variant bioentity names, and fail to identify newly introduced terms that appear as disciplines progress.

State-of-the-art NER systems employ supervised or semi-supervised machine learning. Supervised learning requires pre-annotated gold standard data from which to learn relevant patterns and features for later annotation of previously unseen data. Semi-supervised learning may also use gold standard annotations, but often relies on information contained in lexicons and ontologies to identify entities in the text. Therefore, adapting

[2] http://www.nactem.ac.uk/software/termine/

[3] http://seas.upenn.edu/∼strctlrn/BioTagger/BioTagger.html

NER strategies to a new domain or sub-domain may require the manual creation of gold standard data or manual intervention by an expert to correct the output of automatic NER software. The creation and/or augmentation of lexicons and similar supporting resources is also typically necessary in order to provide domain-specific terminology used in semi-supervised settings.

As an example, consider a researcher investigating recent advances in gene interaction research documented in publications from a *document repository* such as PubMed Central. The researcher will typically *build a corpus* by selecting a set of appropriate texts from the repository, but in order to find the desired information, it is necessary to identify mentions of the entities in which he or she is interested. This demands that the researcher *annotates the corpus* by applying an NER text analysis service to identify potential gene mentions in the data. However, even specialized NER tools (Furrer et al., 2019) for the biomedical domain perform at rates of about 0.56 F1-score, at best. At this point, human intervention is required to *revise the annotations* by correcting mis-identified occurrences of gene names as well as annotating gene names that the tagger missed. A sophisticated annotation editor that learns from the user's activity and proposes new annotations or modifications can significantly increase the speed of the correction process. The revised annotations are then used to *re-train a machine learning algorithm* that can be applied to other, unannotated texts; results are evaluated, and the training texts are corrected anew, where necessary, by the human user. The researcher *repeats* this overall cycle as many times as necessary until a satisfactory result is obtained.

Note that there are two human-in-the-loop cycles here: a tight cycle, where a classifier is trained within the annotation editor itself to assist the user, and a larger cycle where a classifier is separately trained and used to annotate the corpus.

The above describes only one possible scenario using the combined functionalities of *PubAnnotation*, the *LAPPS Grid*, and *INCEpTION* to create texts annotated for biomedical entities. The three platforms are mutually interconnectable, and so it is possible to initiate one's corpus building/annotation activity from within any one of them and move to the others as needed, without the need to explicitly export data from one platform and import it to another or convert formats to enable cross-platform communication. Table 1 summarizes the extent to which each platform supports the various steps in the domain adaptation process and how it can interconnect with the other platforms to address a given step. Figure 1 provides a graphic rendering of possible interactions among the platforms.

4 Conclusion

Our goal is to provide an easy-to-use framework to support mining of biomedical publications and, ultimately, scientific publications, by providing an ecosystem that facilitates the rapid development of corpora annotated for phenomena in specific domains and sub-domains. We accomplish this by leveraging the capabilities of three independently developed systems, rather than attempting to develop a single, monolithic system. While monolithic systems tend to be faster to build and are able to better reflect the needs of a particular use case, their maintenance and long-term sustainability is limited by the attention of their developer community. An approach combining the capabilities of multiple platforms reduces the risk of becoming unmaintained. And, even if one platform becomes unavailable or no longer maintained, making them interoperable inherently requires the development of suitable and generic APIs and data formats, which in turn facilitates connecting with new platforms to replace a lost one or expand the overall ecosystem. For users, this means a reduced risk of being locked in to a particular technology and the ability to pick and combine tools best suited for their task from a wider selection.

Acknowledgements

The LAPPS Grid is supported by US National Science Foundation grants NSF-ACI 1147944, NSF-ACI 1147912, NSF-EAGER 181123, and the Andrew K. Mellon Foundation; the LAPPS Grid is also supported by the NSF-funded Jetstream cloud computing environment (Fischer et al., 2015; Towns et al., 2014).

This work was supported by the German Research Foundation under grant No. EC 503/1-1 and GU 798/21-1 (INCEpTION).

PubAnnotation is supported by the Database Integration Coordination Program funded by National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

References

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Richard Eckart de Castilho, Nancy Ide, Jin-Dong Kim, Jan-Christoph Klie, and Keith Suderman. 2019. Towards cross-platform interoperability for machine-assisted annotation. page to appear. Genomics Inform.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, 9(2):e1002854.

Mahmoud El-Haj, Paul Rayson, Scott Piao, and Jo Knight. 2018. Profiling medical journal articles using a gene ontology semantic tagger. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 4593–4597, Miyazaki, Japan. European Language Resources Association (ELRA).

David Ferrucci, Adam Lally, Karin Verspoor, and Eric Nyberg. 2009. Unstructured information management architecture (UIMA) version 1.0. OASIS Standard.

Jeremy Fischer, Steven Tuecke, Ian Foster, and Craig A. Stewart. 2015. Jetstream: A distributed cloud infrastructure for underresourced higher education communities. In *Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models*, SCREAM '15, pages 53–61, New York, NY, USA. ACM.

Lenz Furrer, Anna Jancso, Nicola Colic, and Fabio Rinaldi. 2019. Oger++: hybrid multi-type entity recognition. *Journal of Cheminformatics*, 11(1):7.

T. Götz and O. Suhre. 2004. Design and implementation of the UIMA common analysis system. *IBM Systems Journal*, 43(3):476 –489.

Marie Hinrichs, Thomas Zastrow, and Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 489–493, Valletta, Malta. European Language Resources Association (ELRA).

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Di Wang, Keith Suderman, Marc Verhagen, and Jonathan Wright. 2014. The Language Applications Grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 22–30, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nancy Ide, Keith Suderman, Marc Verhagen, and James Pustejovsky. 2016. The Language Applications Grid Web Service Exchange Vocabulary. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure - Volume 9442*, WLSI 2015, pages 18–32, New York, NY, USA. Springer-Verlag New York, Inc.

Yang Jin, Ryan T McDonald, Kevin Lerman, Mark A Mandel, Steven Carroll, Mark Y Liberman, Fernando C Pereira, Raymond S Winters, and Peter S White. 2006. Automated recognition of malignancy mentions in biomedical literature. *BMC Bioinformatics*, 7(492).

Jin-Dong Kim and Yue Wang. 2012. Pubannotation - a persistent and sharable corpus and annotation repository. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205, Montréal, Canada. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Penny Labropoulou, Dimitris Galanis, Antonis Lempesis, Mark Greenwood, Petr Knoth, Richard Eckart de Castilho, Stavros Sachtouris, Byron Georgantopoulos, Stefania Martziou, Lucas Anastasiou, Katerina Gkirtzou, Natalia Manola, and Stelios Piperidis. 2018. Openminted: A platform facilitating text mining of scholarly content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, Paris, France. European Language Resources Association (ELRA).

Rafal Rak, Andrew Rowley, Jacob Carter, and Sophia Ananiadou. 2013. Development and analysis of nlp pipelines in argo. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 115–120, Sofia, Bulgaria. Association for Computational Linguistics.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. Xsede: Accelerating scientific discovery. *Computing in Science Engineering*, 16(5):62–74.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics (PCI 2005), LNCS 3746*, pages 382–392.

Marc Verhagen, Keith Suderman, Di Wang, Nancy Ide, Chunqi Shi, Jonathan Wright, and James Pustejovsky. 2016. The LAPPS Interchange Format. In *Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure - Volume 9442*, WLSI 2015, pages 33–47, New York, NY, USA. Springer-Verlag New York, Inc.