Mining Biomedical Publications With The LAPPS Grid

Nancy Ide*, Keith Suderman*, and Jin-Dong Kim**
*Department of Computer Science, Vassar College
Poughkeepsie, New York USA
{ide, suderman}@cs.vassar.edu
*Database Center for Life Science (DBCLS), Univ. of Tokyo Kashiwa-no-ha
Wakashiba, Kashiwa-shi, Chiba, Japan
jdkim@dbcls.rois.ac.jp

Abstract

It is widely recognized that the ability to exploit Natural Language Processing (NLP) text mining strategies has the potential to increase productivity and innovation in the sciences by orders of magnitude, by enabling scientists to pull information from research articles in scientific disciplines such as genomics and biomedicine. The Language Applications (LAPPS) Grid is an infrastructure for rapid development of natural language processing applications (NLP) that provides an ideal platform to support mining scientific literature. Its Galaxy interface and the interoperability among tools together provide an intuitive and easy-to-use platform, and users can experiment with and exploit NLP tools and resources without the need to determine which are suited to a particular task, and without the need for significant computer expertise. The LAPPS Grid has collaborated with the developers of PubAnnotation to integrate the services and resources provided by each in order to greatly enhance the user's ability to annotate scientific publications and share the results. This poster/demo shows how the LAPPS Grid can facilitate mining scientific publications, including identification and extraction of relevant entities, relations, and events; iterative manual correction and evaluation of automatically-produced annotations, and customization of supporting resources to accommodate specific domains.

Keywords: biomedical text mining, LAPPS Grid, PubAnnotation, Open Advancement evaluation, reproducibility

1. Introduction

Keeping up with the ever-expanding flow of data and publications is untenable and poses a fundamental bottleneck to scientific progress. The global research community generates approximately 2.5 million new scholarly papers per year (in English only); a new research paper is published every 12 seconds. Current search technologies typically find many relevant documents, but they do not extract and organize the information content of these documents or suggest new scientific hypotheses based on this organized content.

It is widely recognized that the ability to exploit Natural Language Processing (NLP) text mining strategies has the potential to increase productivity and innovation in the sciences by orders of magnitude, by enabling scientists to pull information from research articles in scientific disciplines such as genomics and biomedicine. The application of NLP techniques can also lead to hypotheses and discoveries for which there is "hidden" (not explicitly stated) evidence in the research literature and enable linking extracted information to form new facts or new hypotheses to be explored further. These methods enable scientists to rapidly identify publications relevant to their own research as well as make scientific discoveries by scouring hundreds of research papers for associations and connections (such as between drugs and side effects, or genes and disease pathways) that humans reading each paper individually might not notice.

Up to now, the use of NLP technologies has required considerable skill in the field. However, recent development of environments for constructing customizable NLP applications has opened the door for scientists to exploit NLP technologies for discovering and mining information from massive bodies of scientific publications such as those found in PubMed, PLoS, Web of Science, etc.

The Language Applications (LAPPS) Grid[1] (Ide et al., 2014) provides an infrastructure for rapid development of natural language processing applications (NLP) by providing access to a wide range of tools and making them both syntactically and semantically interoperable. The LAPPS Grid uses the Galaxy platform[2] (Giardine et al., 2005), originally developed for use by genomics researchers with little computational expertise, as its workflow engine. The Galaxy interface and the interoperability among tools together provide an intuitive and easy-to-use platform that enables users to experiment with and exploit NLP tools and resources without the need to determine which are suited to a particular task, and without the need for significant computer expertise.

We demonstrate how the LAPPS Grid can be used to rapidly and easily develop out-of-the-box workflows for information and relation extraction and adapt them to data for specific disciplines, for example by providing means to rapidly bootstrap custom dictionaries and gazetteers. We also show how users can employ a cycle of automatic annotation and manual correction to create more robust annotations, and exploit state-of-the-art evaluation services to determine the optimal tool and resource configuration for a given task. Finally, we demonstrate the use of the LAPPS Grid access major scientific publications databases stored in the cloud as well as materials and facilities available through the PubAnnotation portal[3], and query them with Apache Solr for data discovery and mining.

[1] http://www.lappsgrid.org
[2] http://galaxyproject.org
[3] http://pubannotation.org

2. LAPPS Grid Overview

The US National Science Foundation SI^2-funded Language Applications (LAPPS) Grid[4] (Ide et al., 2014) was originally developed to facilitate rapid development of Natural Language Processing (NLP) applications. From the outset, it was designed to enable sophisticated analyses while hiding the complexities associated with the underlying infrastructure, with an eye toward serving the needs of researchers and students who may not have substantial computational expertise. The LAPPS Grid provides seamless access to a wide range of NLP tools, including popular public tools such as StanfordNLP, OpenNLP, NLTK, LingPipe, as well as tools and modules available in GATE and various UIMA platforms; machine learning facilities; and a state-of-the-art Open Advancement (OA) evaluation system developed at Carnegie Mellon University and used in the development of IBM's Jeopardy-winning Watson. It also provides access to several mainstream resources, including holdings of the Linguistic Data Consortium (LDC). Most crucially, the LAPPS Grid provides for using all of these tools and resources *interoperably* in a seamless "plug-and-play" workflow environment, thereby eliminating the effort required to harmonize input and output formats to use a set of tools together. The LAPPS Grid is also flexible and extensible, as tools and datasources are routinely added to the LAPPS Grid as required by various researchers and projects.

The LAPPS Grid is open source (Apache 2.0 license) and free for use by anyone, and can be run from the web, on a user's laptop or desktop, in the cloud, or as a self-contained docker image when it is necessary to protect sensitive data or no network connection is available. We have also developed means to provide secure access to licensed data and software where necessary from within the LAPPS Grid, as well as to allow for user authentication and identification through identity providers (e.g., InCommon[5]) that provide a secure and privacy-preserving trust fabric for their members.

The LAPPS Grid provides cloud-based computation via the NSF-funded Extreme Science and Engineering Discovery Environment (XSEDE)[6] and the associated Jetstream[7] cloud environment. These resources allow users to create virtual machines configured as specialized versions of the LAPPS Grid on the remote resource. If necessary, access can be given to specified domain servers holding secure data.

2.1. Galaxy workflow engine

The LAPPS Grid's adaptation of the Galaxy workflow engine provides an intuitive and easy-to-use interface and data management system. The Galaxy project[8] started in 2005 to create a system enabling biologists without informatics expertise to perform computational analysis through the web. It has since been widely adopted within the life sciences community.

Galaxy is an open-source application[9] that includes tool integration and history capabilities together with a workflow system for building automated multi-step analyses, a visualization framework including visual analysis capabilities, and facilities for sharing and publishing analyses (Goecks et al., 2010; Afgan et al., 2016). It is accessed through a graphical interface where data inputs and computational steps are selected from dynamic menus, and results are displayed in plots and summaries that encourage interactive workflows and the exploration of hypotheses.

With funding from the National Science Foundation[10], we are working with the Galaxy development team in order to adapt the system to our domain and apply Galaxy's powerful analytic and visualization software to information extracted from texts using the LAPPS Grid without leaving the platform. The ultimate goal of this collaboration is to both enhance the capabilities required to support NLP application development and contribute to the expansion of Galaxy to domains outside the life sciences, which is a current goal of the Galaxy project.

2.2. Comparison to existing platforms for biomedical text analysis

The two most well-known platforms that currently support scientific literature mining are the UIMA-based U-Compare (Kano et al., 2008) and a more recently developed platform named Argo (Rak et al., 2012). Both of these systems allow the user to assemble modular pipelines and perform evaluations against a gold standard. U-Compare is plagued by instabilities of platform interoperability, permissions, and the like, and typically requires the intercession of a specialized software engineer. Argo attempts to ameliorate some of these problems by providing a web-based interface to the underlying UIMA-based system, but suffers from many of the same problems as U-Compare and is seemingly unsupported at this time. Frameworks that support general text mining, e.g., the General Architecture for Text Engineering (GATE) (Cunningham et al., 2011), provide "local interoperability" for tools available from within the framework, but there is no interoperability with tools or components available from outside the framework that the user might wish to use. In contrast, the LAPPS Grid provides interoperable access to tools in various UIMA systems as well as tools from GATE, which can be pipelined within the LAPPS Grid without the need for I/O format conversion.[11].

2.3. Access to Resources

Access to publication resources in the biomedical domain is problematic for several reasons:

1. Repositories vary in the types and format of their contents, some containing abstracts while others contain full text articles. Many provide only access to query

[4] http://www.lappsgrid.org

[5] http://incommon.org

[6] https://www.xsede.org

[7] https://jetstream-cloud.org

[8] http://galaxyproject.org

[9] Distributed under the terms of permissive Academic Free License: http://getgalaxy.org

[10] US NSF grant ABI 1661497

[11] The LAPPS Grid currently contains several GATE tools and makes them interoperable with all other tools in the framework, including several UIMA-based tools from DKPro.

results and not the full text itself. For those that do deliver the full text, many provide them in PDF format, which requires sophisticated conversion to enable text mining, while others deliver XML, JSON, and a variety of other formats.

2. Repositories are spread out all over the web, making them difficult to find and even harder to use together.

3. While some repositories are freely open, others require subscriptions or licenses.

An instance of the LAPPS Grid tailored to mining biomedical publications is currently maintained on the JetStream[12] cloud environment.[13] The instance currently includes full-text PubMed data used in several BioNLP and SemEval shared tasks, some with several layers of annotation. In addition, the LAPPS Grid has collaborated with the developers of PubAnnotation[14] to integrate the services and resources provided by each in order to greatly enhance the user's ability to annotate scientific publications and share the results. PubAnnotation is a repository of text annotations applied to biomedical publications, all of which are aligned to the canonical text in either PubMed or PubMed Central, thus linking all PubAnnotation annotations to each other through the canonical texts. Annotations are accessible and searchable through standard web protocols such as the REST API. Through the collaboration with PubAnnotation we have enabled access to holdings of PubMed (ca. 12 million abstracts) and PubMed Central (over 11 thousand full-text documents) by creating a "PubAnnotation datasource" in the LAPPS Grid. Users also have access to the annotations in the PubAnnotation repository that are linked to the texts.

Note that all text resources in the LAPPS Grid are rendered in the JSON-LD-based LAPPS Interchange Format (LIF) (Verhagen et al., 2015), which means that they are consumable by all tools in the LAPPS Grid.

We aim to provide access to as many full text articles as possible from within the LAPPS Grid as direct datasources, thus providing a "one stop" location for accessing publications available from otherwise scattered locations. The articles are converted to our internal JSON-LD format for delivery within the Grid so that researchers need not be concerned with issues of format conversion. For the purposes of mining scientific publications, we are interested in repositories that deliver full text, such as the following: PubMed Central (PMC)[15] (open section), PubMed Central Canada[16], Open Science Repository[17], Public Library of Science (PLOS)[18], arXiv[19], and BioMed Central[20]. We also plan to provide access for users with appropriate credentials to data from repositories that require a subscription, such as the Web of Science[21], using the authentication procedures already in place for delivering data from the Linguistic Data Consortium (LDC) within the Grid.

The LAPPS Grid also integrates several existing lexicons, ontologies, and knowledge bases relevant to the fields in which our collaborators are working (e.g., UMLS[22], EntrezGene[23], MeSH[24], UniProt[25]) as datasources, thus eliminating the need to access each one from a different source. In addition, Galaxy itself provides access to a very wide range of datasources relevant to genomic analysis and gene sequencing, as well as multiple tools to convert and manipulate the data.

3. Tools for Biomedical Text Analysis

In addition to the wide range of general purpose NLP tools available in the web-based LAPPS Grid instance[26], the JetStream instance currently includes several entity recognizers for biomedical terminology, event annotators, relation extraction software, and dictionary-based entity recognizers that can use customized lexicons, as well as machine learning facilities and tools for identifying key terms, synonyms, acronyms, lexical variants, and the like.

Figure 1 shows a portion of the LAPPS Grid menu in Galaxy that includes biomedical annotation tools together with a visualization of protein annotation on a PubMed document, using the Brat visualization tool included in the Grid.

4. Open Advancement Evaluation

One of the most valuable components of the LAPPS Grid suite of services for publication mining is its OA evaluation capability. Most information systems consist of a number of processing units or components arranged in series, or workflow; OA enables the user to automatically and efficiently evaluate different workflow configurations and identify those that achieve the best results. Much current research focuses on experimentation with parameters of a single module while keeping modules and parameters elsewhere in the system frozen. For example, a typical, simple workflow for biomedical text mining will rely on algorithms, toolkits, and pre-trained models for basic NLP processes such as sentence segmentation, tokenization, part-of-speech tagging, and chunking, coupled with tools and resources specially suited to a given area of biomedical research that extract terms and entities, identify terms and acronyms, provide synonyms and lexical variants, etc. Improvement of the system's performance might focus on refining entity extraction, possibly applying different models and/or augmenting a gazetteer or lexicon, while the preliminary processes remain static. There is typically no knowledge of the degree to which the performance of any individual module contributes to overall performance, although

[12] https://jetstream-cloud.org

[13] This instance can be accessed at http://jetstream.lappsgrid.org.

[14] http://www.pubannotation.org

[15] https://www.ncbi.nlm.nih.gov/pmc/

[16] http://pubmedcentralcanada.ca/pmcc/

[17] http://www.open-science-repository.com

[18] https://www.plos.org

[19] https://arxiv.org/

[20] https://www.biomedcentral.com

[21] https://clarivate.com/products/web-of-science/

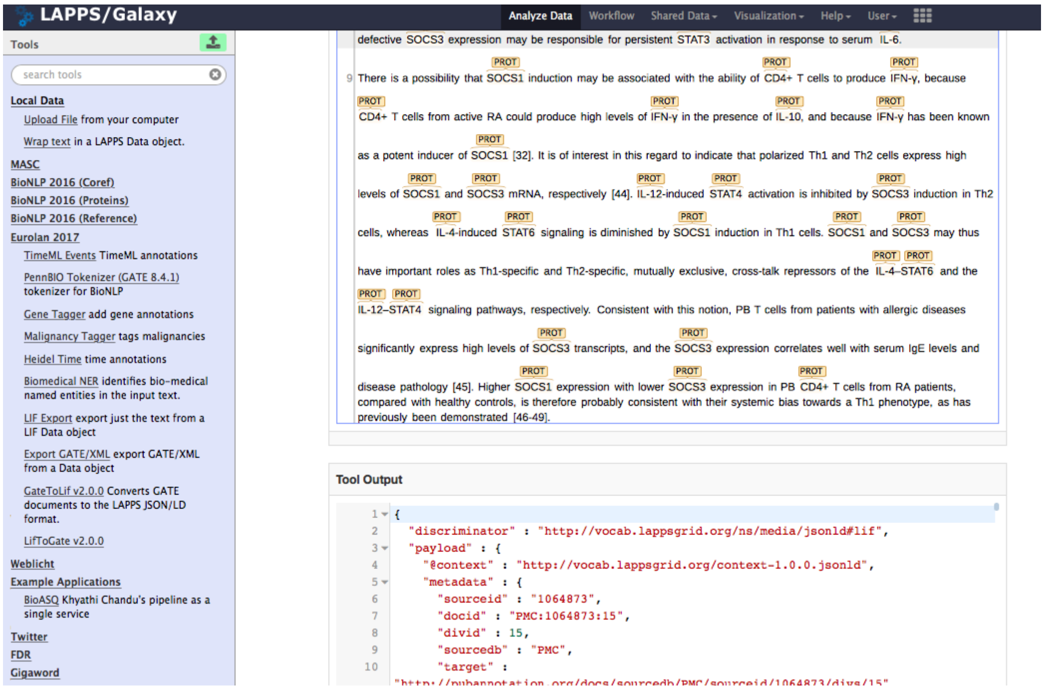
[22] https://www.nlm.nih.gov/research/umls/

[23] http://www.ncbi.nlm.nih.gov/gene

[24] https://www.nlm.nih.gov/mesh/

[25] http://www.uniprot.org

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



Local Data

Upload File from your computer

Wrap text in a LAPPS Data object.

MASC

BioNLP 2016 (Coref)

BioNLP 2016 (Proteins)

BioNLP 2016 (Reference)

Eurolan 2017

TimeML Events TimeML annotations

PennBIO Tokenizer (GATE 8.4.1) tokenizer for BioNLP

Gene Tagger add gene annotations

Malignancy Tagger tags malignancies

Heidel Time time annotations

Biomedical NER identifies bio-medical named entities in the input text.

LIF Export export just the text from a LIF Data object

Export GATE/XML export GATE/XML from a Data object

GateToLif v2.0.0 Converts GATE documents to the LAPPS JSON/LD format.

LifToGate v2.0.0

Weblicht

Example Applications

BioASQ Khyathi Chandu's pipeline as a single service

Twitter

FDR

Gigaword

defective SOCS3 expression may be responsible for persistent STAT3 activation in response to serum IL-6.

9 There is a possibility that SOCS1 induction may be associated with the ability of CD4+ T cells to produce IFN-γ, because CD4+ T cells from active RA could produce high levels of IFN-γ in the presence of IL-10, and because IFN-γ has been known as a potent inducer of SOCS1 [32]. It is of interest in this regard to indicate that polarized Th1 and Th2 cells express high levels of SOCS1 and SOCS3 mRNA, respectively [44]. IL-12-induced STAT4 activation is inhibited by SOCS3 induction in Th2 cells, whereas IL-4-induced STAT6 signaling is diminished by SOCS1 induction in Th1 cells. SOCS1 and SOCS3 may thus have important roles as Th1-specific and Th2-specific, mutually exclusive, cross-talk repressors of the IL-4–STAT6 and the IL-12–STAT4 signaling pathways, respectively. Consistent with this notion, PB T cells from patients with allergic diseases significantly express high levels of SOCS3 transcripts, and the SOCS3 expression correlates well with serum IgE levels and disease pathology [45]. Higher SOCS1 expression with lower SOCS3 expression in PB CD4+ T cells from RA patients, compared with healthy controls, is therefore probably consistent with their systemic bias towards a Th1 phenotype, as has previously been demonstrated [46-49].

Tool Output

```
{
  "discriminator" : "http://vocab.lappsgrid.org/ns/media/jsonld#lif",
  "payload" : {
    "@context" : "http://vocab.lappsgrid.org/context-1.0.0.jsonld",
    "metadata" : {
      "sourceid" : "1064873",
      "docid" : "PMC:1064873:15",
      "divid" : 15,
      "sourcedb" : "PMC",
      "target" :
"http://pubannotation.org/docs/sourcedb/PMC/sourceid/1064873/divs/15"
```

Figure 1: Some biomedical annotation tools and a visualization of protein annotation in the LAPPS Grid

earlier tools in the pipeline could in fact contribute significantly to poorer than expected results. The OA evaluation strategy overcomes this by exposing the performance of individual modules along with overall performance; an easy-to-use framework for building and testing different tool configurations such as the LAPPS Grid makes it easy to examine alternative tool chains and determine the optimal configuration.

As an example, (Yang et al., 2013) applied the OA approach to implement biomedical information systems for question answering tasks from the TREC Genomics. The OA framework automatically evaluated different system configurations and identified those that achieved better results than prior published results. The study found some simple and relatively unexplored contributors to improved performance including, for example, leveraging various sources varied for synonym expansion and acronym expansion, and altering weights for concept terms and verbs (concept terms favored higher weights and verbs favored moderate weights). Although automatic search through alternative pipelines is not yet implemented in the LAPPS Grid[27], even manual iteration over different configurations based on detailed information about each module's contribution can both improve results and reveal the impacts of processing components that are typically overlooked, due to the LAPPS Grid's wide range of modules for various tasks and the ease of workflow construction and modification.

5. Integration of Annotation Facilities

PubAnnotation includes TextAE, a powerful and easy-to-use Javascript app for text annotation and visualization. In our collaboration with PubAnnotation, we have enabled LAPPS Grid users to invoke TextAE from within the Grid, and we have similarly enabled PubAnnotation users to access and apply LAPPS Grid tools from within the PubAnnotation environment. Reciprocal access between PubAnnotation and the LAPPS Grid means that users can easily apply automatic annotation tools and subsequently manually correct annotations, in an iterative "human-in-the-loop" process of refinement. This is especially useful for the creation of training data for machine learning, especially iterative, semi-supervised approaches such as active learning. This, coupled with the Open Advancement evaluation facilities, provides a powerful environment for rapid development of high-quality automatic annotation procedures.

The LAPPS Grid provides means for users to easily register and thus share annotations in the PubAnnotation registry. As noted above, annotations registered in PubAnnotation are aligned with the canonical text and all other annotations applied to the same data. Conversely, annotations from the PubAnnotation registry, or annotations created by users from within the PubAnnotation platform using the TextAE editor, can be imported into the LAPPS Grid for further automatic processing, for example, application of tools that use these annotations in order to produce additional annotation layers.

Data and annotations in PubAnnotation are stored in a JSON format[28]; LAPPS Grid services use the Grid's JSON-LD format (LIF). For communication between the two platforms, we automatically convert between the two formats so that interoperability is seamless from the point of view of the user. Conversion has dictated some minor modifications to the PubAnnotation format, including the addition of metadata that is required by LAPPS Grid services, but is

[27] Search over multiple configuration spaces in the LAPPS Grid OA component is currently under development.

[28] http://www.pubannotation.org/docs/annotation-format/

otherwise relatively straightforward.

6. Support for Replicability and Reproducibility of Results and Data Sharing

There is an ongoing concern in the biomedical community concerning the transparency and reproducibility of published research (Ioannidis, 2005; Iqbal et al., 2016). Our adaptation of the Galaxy workflow system fosters replicability and reproducibility for biomedical text mining studies by providing the following capabilities[29]:

- automatic recording of inputs, tools, parameters and settings used for each step in an analysis in a publicly viewable history, thereby ensuring that each result can be exactly reproduced and reviewed later;
- provisions for sharing datasets, histories, and workflows via web links, with progressive levels of sharing including the ability to publish in a public repository;
- ability to create custom web-based documents to communicate about an entire experiment, which takes a step towards the next generation of online publications that would include both full paper and all supporting materials.

In addition to enabling other users to replicate an experiment, the individual user can develop a rich, organized catalog of reusable workflows rather than starting from scratch each time or trying to navigate a collection of *ad hoc* analysis scripts. Similarly, it is possible to repeatedly apply a command history on different data. Once an analysis is done, the record eliminates ambiguity as to which result used which settings and provides critical information for follow-up analysis.

Within the LAPPS Grid and its community of users, sharing of newly created resources (lexicons, etc.) for use by others from within the Grid, as well as for the purposes of replicability and reproducibility, is strongly encouraged. We anticipate that Grid users will increasingly create new resources of this kind, especially as we develop better facilities for domain adaptation (see Section 9.).

7. Support for Data Privacy

Privacy constraints often make it necessary to protect biomedical and clinical data from exposure to network access. In addition, research activity can be sensitive or proprietary and require protection from outside access. To address this need, a docker image containing a self-enclosed instance of the LAPPS Grid can be installed locally on a user's machine or server and used to access and process local data, thus disabling access via the internet. A local docker instance can also be used when a network connection is not available for any reason.

In addition to ensuring privacy, creation of a docker image containing a particular instance of the LAPPS Grid can provide absolute replicability for results, by encapsulating specific versions of tools, parameters, workflows, and data used in an experiment; any person wishing to replicate the original results or apply the methods to new data can easily do so without attempting to recreate the original environment, and with assurance that that environment is exactly as reported.

8. Handling Large-scale Data

An issue to be dealt with in text mining of large publication databases is the ability to handle high-throughput data at scale. The LAPPS Grid uses XSEDE, a heterogeneous high performance computing (HPC) environment for research, and the associated Jetstream cloud environment, for large-scale analyses and storage. The Grid will benefit tremendously in the near future from the recently-funded NSF ABI project (NSF ABI 1661497) that is linking Galaxy and XSEDE in order to provide HPC processing capabilities and massive storage through the Galaxy platform. To enable fast query of publication databases, we plan to create full text indexes with Apache Solr[30] of open scientific publication databases (PubMed, PLoS, Web of Science, etc.) for distributed indexing and load-balanced querying. The indexes will be stored in the cloud on Jetstream. We will also regularly (e.g., weekly) regenerate the Solr indexes so that the most recent material is available.

9. Future Work: Support for Domain Adaptation

Different domains and individual researchers have specific goals and knowledge interests, but it is all too often the case that an available system doesn't target the information of specific interest and is therefore not useful. We see development of methods for domain adaptation for text mining as a critical, and *as yet unaddressed*, need for mining scientific publications.

Scientific publications share commonalities of structure and style, but across domains and areas of specialization they differ most drastically in the use of highly specialized vocabularies and terminology, which may comprise as much as 12% of overall document vocabulary. Domain adaptation for scientific publications therefore necessarily focuses on handling previously unseen vocabulary and terms. Various vocabularies, ontologies, and knowledge bases exist in the field, but these resources cover only a fraction of the vocabulary, especially for specialized subdomains. Furthermore, the constant introduction of new terms and short forms or abbreviations makes vocabulary adaptation for scientific publications an ongoing activity.

Domain adaptation for scientific text mining therefore involves the ability to modify and extend existing lexicons and other supporting resources. To do this, the scientist must be able to examine results, identify unrecognized terms and false positives, retrain a recognition module using the new information, and run the workflow again. Thus domain adaptation involves an iterative cycle of performance improvement (active learning).

[29] See (Goecks et al., 2010) for a comprehensive overview of Galaxy's sharing and publication capabilities, and (Sandve et al., 2013) for further discussion.

[30] http://lucene.apache.org/solr/

In Section 5. we briefly outlined how a user might exploit within the LAPPS Grid/PubAnnotation integration to facilitate the development of training data; a similar iterative process can be applied to development of supporting resources such as lexicons and term banks. We see domain adaptation as an increasingly important need for the future development of publication mining capabilities; by providing facilities to support this activity, as well as supporting and encouraging sharing of resources and methods, we hope to move the field forward.

10. Conclusion

Support for biomedical text mining is now a major focus for LAPPS Grid development. In addition to the facilities and collaborations with the Galaxy and PubAnnotation projects described above, we have just begun working on mining and summarizing clinical reports in a collaboration with the US Center for Disease Control (CDC) and Food and Drug Association (FDA), which will extend our work to a broader range of text types and applications.

At this time, tools in the LAPPS Grid focus on English. However, in a project funded by the A. K. Mellon Foundation, we are establishing an interoperable "bridge" with the CLARIN WebLicht framework (Dima et al., 2012) hosted by the CLARIN-D Center in Tübingen, Germany[31], which will give LAPPS Grid users access to a vast range of NLP tools and data available from WebLicht and CLARIN Centers throughout Europe. Access to tools for multi-lingual analyses will extend LAPPS Grid capabilities to multiple languages. In addition, because the LAPPS Grid is federated with seven other grids in the Federated Grid of Language Services (FGLS) (Ishida et al., 2014), including the Language Grid housed at Kyoto University[32], users will have interoperable access to atomic and composite web services for Asian languages available from any of these federated grids.

As a final note, the LAPPS Grid, Galaxy, and PubAnnotation are all open source projects that invite contributions from developers and users. In particular, the LAPPS Grid seeks the contribution of tools and resources for biomedical text mining to augment the current facilities. Thus we hope to build up a truly useful platform that is available to everyone, whatever their goals.

11. Acknowledgements

This work is supported by National Science Foundation grants NSF-ACI 1147944, NSF-ACI 1147912, and NSF ABI 1661497.

12. Bibliographical References

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Kuster, G. V., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., and Goecks, J. (2016). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(Webserver-Issue):W3–W10.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Dima, E., Hinrichs, E., Hinrichs, M., Kislev, A., Trippel, T., and Zastrow, T. (2012). Integration of weblicht into the clarin infrastructure. In *Proceedings of the Joint CLARIN-D/DARIAH Workshop at Digital Humanities Conference 2012: Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts*, pages 17–23, Hamburg, Germany.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451–55.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11:R86.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., and Wright, J. (2014). The language application grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8).

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., and Ioannidis, J. P. A. (2016). Reproducible Research Practices and Transparency across the Biomedical Literature. *PLOS Biology*, 14(1).

Ishida, T., Murakami, Y., Lin, D., Nakaguchi, T., and Otani, M. (2014). Open Language Grid–Towards a Global Language Service Infrastructure. In *The Third ASE International Conference on Social Informatics (SocialInformatics 2014)*, Cambridge, Massachusetts, USA.

Kano, Y., Nguyen, N., Sætre, R., Yoshida, K., Fukamachi, K., Miyao, Y., Tsuruoka, Y., Ananiadou, S., and Tsujii, J. (2008). Towards Data And Goal Oriented Analysis: Tool Inter-Operability And Combinatorial Comparison. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 859–864, Hyderabad, India, January.

Rak, R., Rowley, A., Black, W., and Ananiadou, S. (2012). Argo: An integrative, interactive, text mining-based workbench supporting curation. *Database*, 2012:bas010.

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10).

Verhagen, M., Suderman, K., Wang, D., Ide, N., Shi, C., Wright, J., and Pustejovsky, J. (2015). The LAPPS Interchange Format. In *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure WLSI (2015)*, pages 33–47, Kyoto,

[31] This project is described in a separate submission to LREC 2018.

[32] http://langrid.org

Japan. Springer International Publishing.

Yang, Z., Garduno, E., Fang, Y., Maiberg, A., McCormack, C., and Nyberg, E. (2013). Building Optimal Information Systems Automatically: Configuration Space Exploration for Biomedical Information Systems. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1421–1430, New York, NY, USA. ACM.