Demonstration: The Language Application Grid as a Platform for Digital Humanities Research

Nancy Ide, Keith Suderman
Department of Computer Science
Vassar College
E-mail: {ide,suderman}@cs.vassar.edu

James Pustejovsky
Department of Computer Science
Brandeis University
E-mail: jamesp@cs.brandeis.edu

Abstract

The LAPPS Grid project, which has developed a platform providing access to a vast array of language processing tools and resources for the purposes of research and development in natural language processing (NLP), has recently expanded to enhance its usability by non-technical users such as those in the DH community. We provide a live demonstration of LAPPS Grid use, ranging from "from scratch" construction of a workflow using atomic tools to a pre-configured docker image that can be run off-the-shelf on a laptop or in the cloud, for several tasks of relevance to the DH community.

1 Introduction

Over the past few years, Digital Humanities (DH) has looked to Computational Linguistics (CL) for methods to enable richer analysis of literary, historical, and other kinds of documents, recognizing that CL methods and procedures can in fact enhance the kinds and amount of information that can be automatically extracted from language data [14]. However, several obstacles have prevented humanists from wholesale adoption of CL tools, the most well known of which is that they are typically difficult to use without a fair amount of technical background. Other, more subtle but perhaps more deeply rooted obstacles have also contributed, most notably dramatic differences in perspective, approach, and simply differences in the language data that each community typically deals with. It is only recently that CL methods and tools have begun to be made more accessible to non-technical users and are beginning to be widely adopted by the DH community; however, there remains considerable work to be done to fully adapt CL tools and methods to use by DH scholars.

The Language Applications (LAPPS) Grid [6] is an NSF-funded project involving Vassar College, Brandeis University, Carnegie Mellon University, and the Linguistic Data Consortium at the University of Pennsylvania. The original motivation for the project, begun in 2012, was to address the endemic lack of interoperability among CL tools and data that has plagued the CL field for decades. Atomic natural language processing (NLP) tools (e.g., part of speech taggers, syntactic analyzers, entity detectors, etc.) are typically *pipelined* to create more sophisticated applications; the lack of interoperability among tools, corpora, and other language resources often leads to considerable waste of effort to make them work together in a pipeline, or "workflow". To overcome the problem, the LAPPS Grid project undertook to engineer a platform that both provides access to a wide array of language processing tools and resources, and exploits recognized standards and best practices to negotiate incompatibilities for the user.

Over the past five years the LAPPS Grid project has collaborated with several major projects in the US, Europe, and Asia to expand its range of accessible tools and resources as well as to augment the capabilities of the platform. Our collaborators serve a broad range of users, well beyond the NLP community we originally intended to serve, including users involved in inter-cultural communication and users from the DH community. We have also begun to create purpose-built instances of the LAPPS Grid to use in courses aimed at non-technical users, and we are currently working with a major project in the digital humanities [12] and pursuing funding to collaborate with several others. As a result, the LAPPS Grid is continually increasing its usability by non-technical users such as those in the DH community.

Our demonstration provides several sample usages of the LAPPS Grid relevant to digital humanities research, including tool pipelines developed "from scratch" as well as pre-configured workflows that can be used as is, and demonstrates both the analysis and creation of resources.

2 LAPPS Grid Overview

The LAPPS Grid is an open platform that provides access to hundreds of NLP tools and language resources. It incorporates the Galaxy workflow and data management framework [5], which was developed by researchers in the field of genomics and specifically designed to enable researchers in the life sciences to access resources and compose applications without requiring technical expertise. The LAPPS Grid is very flexible and configurable: it can be accessed through a web interface (http://galaxy.lappsgrid.org), deployed locally on any Unix system (laptop, desktop, or server), or run from the cloud. Another feature of the LAPPS Grid is its Open Advancement (OA) Evaluation system, which enables the user to explore variant pipelines involving alternative tools in order to identify the most effective configuration in terms of precision and recall.

The LAPPS Grid is part of the Federated Grid of Language Services (FGLS)

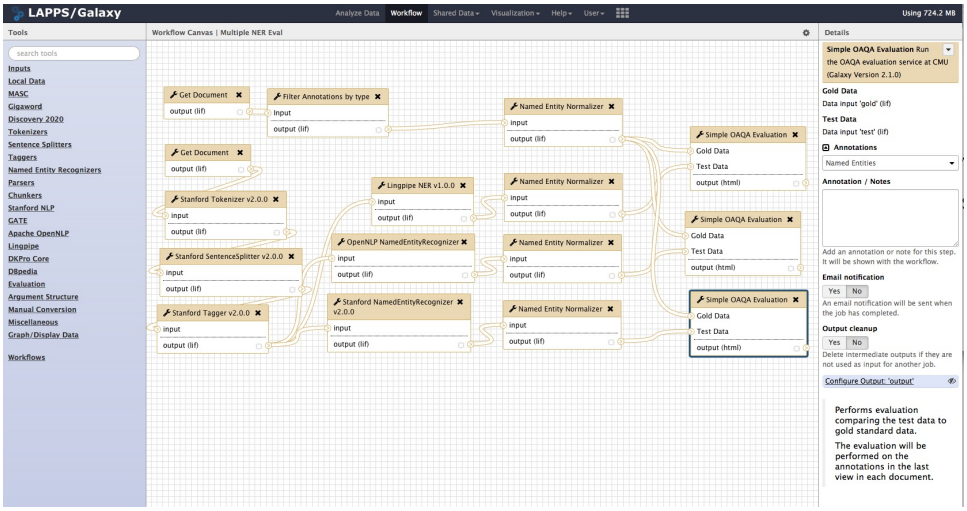

Figure 1: Workflow evaluating three entity recognizers.

[7], an international network of grids including the University of Kyoto's Language Grid[1] and several other Asian and European grids. We have recently entered into a Mellon-funded federation with the pan-European CLARIN project's WebLicht/Tübingen[2] and LINDAT/CLARIN (Prague)[3] frameworks, whose focus is to provide support for humanities and social science scholarship. These two collaborations provide seamless access to all of the tools and resources in any one of the federated platforms for the LAPPS Grid user. Thus we have vastly increased the availability of multi-lingual and multi-modal resources and tools in the LAPPS Grid, and, through our collaboration with CLARIN, expanded the range of services applicable to DH research.

3 CL for DH in the LAPPS Grid

The LAPPS Grid in its current form addresses many of the needs for DH research. It provides easy-to-use access to a wide variety of customizable low-level CL tools, including tokenizers, sentence boundary detectors, part-of-speech taggers, named entity recognizers, co-referencers, phrase-structure and dependency parsers, among others. It also provides facilities for comparing the effectiveness of tools that perform the same task in order to identify the one that is best suited to the task. For example, Figure 1 shows an evaluation pipeline in Galaxy that compares the output of three named entity recognizers to from gold standard annotations; this example shows each small step in the workflow, but sub-steps (for

[1] http://langrid.org
[2] http://weblicht.sfs.uni-tuebingen.de/
[3] https://lindat.mff.cuni.cz/

example, the Tokenization-SentenceSplitter-Tagger sequence that feeds the three entity recognizers) could be bundled into a workflow and plugged in as a single step.

The datasets used in DH research are diverse, often involving ancient text, texts in languages typically not covered in NLP such as Latin, poetry, historical documents, and multi-media, and in some cases need to be representative across multiple genres. Large CL datasets, on the other hand, are typically largely composed of genres such as newswire (Penn Treebank [9], English Gigaword [10], etc.), or they suffer from problems such as the inclusion of digitization artifacts, opaque and unbalanced sampling, etc. [11, 8]. As a result, readily available NLP tools often perform quite badly on DH data, due to dramatic differences in terminology and entities, syntactic structure, etc. This often necessitates augmenting lexicons, gazetteers, and pattern-matching rules used by these tools for the purposes of DH research. Recent examples include augmentation of a contemporary affective lexicon in order to study affect change patterns in German historical texts between 1740 and 1900 [2], and applying automatic parsing as a “pre-annotation tool" for manual annotation of syntax in Old East Slavic texts [4]. In the LAPPS Grid, these tasks are accomplished by using “human-in-the-loop" capabilities to to perform manual annotation and/or augment existing resources incrementally as new entries or patterns emerge from analysis, without leaving the environment to use external tools. More sophisticated analyses can exploit a cycle of automatic annotation using machine learning followed by manual correction, which can then be used to iteratively enhance the performance of the learning algorithm.

Data visualization is often essential for humanities research, and the LAPPS Grid includes a wide range of statistical and visualization tools. A basic but common task is to generate frequency distributions or distributions across a text, collection, timeline, etc. for any type of phenomenon. For example, one recent study examined the appearance of neologisms and words that become obsolete over several decades of Dutch magazine texts as well as tweets, by generating graphs showing initial and final word frequencies over time intervals [13]. Other projects use visualization of relations in graph form. For example, one study used named entity recognition and co-reference tools to identify characters in the novels comprising *A Song of Ice and Fire* and then generated a weighted graph depicting social relations among characters based on dialogue interactions [15]; while another extracted a dictionary of concepts by parsing the English sentences from multiple translations of Wittgenstein’s *Tractatus Logico-Philosophicus* and inferred semantic relations between concepts using word contexts, eventually generating a graph of inter-relations among concepts [1].

4 Conclusion

The LAPPS Grid demonstration will show how it can be used to perform tasks relevant to DH research such as those described above, as well as many others.

Facilities suitable for DH scholarship and research not currently available in the LAPPS Grid are being regularly added to the platform as we receive input from the DH community, and our current collaboration with the CLARIN projects in Europe will significantly enhance LAPPS Grid facilities for DH research in the near future. In the meantime, LAPPS Grid users already have access to the wide range of tools and resources available through the Language Grid and other federated grids, which focus on machine translation and other facilities for cultural collaboration. A new collaboration with the Alveo project [3] in Australia will provide access to a large suite of tools for analysis of multi-modal data, including video, audio, transcriptions of audio, and tools for their analysis. Ultimately, the LAPPS Grid aims to provide an ever-increasing set of tools for DH research, enhance ease of use for non-technical users, and in general help to move DH toward more empirically-grounded (and replicable) methods.

References

[1] Anca Bucur and Sergiu Nisioi. A Visual Representation of Wittgenstein's Tractatus Logico-Philosophicus. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 71–75, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[2] Sven Buechel, Johannes Hellrich, and Udo Hahn. Feelings from the Past–Adapting Affective Lexicons for Historical Emotion Analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 54–61, Osaka, Japan, 2016. COLING 2016 Organizing Committee.

[3] Steve Cassidy, Dominique Estival, Timothy Jones, Denis Burnham, and Jared Burghold. The Alveo Virtual Laboratory: A Web based Repository API. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[4] Hanne Martine Eckhoff and Aleksandrs Berdicevskis. Automatic parsing as an efficient pre-annotation tool for historical texts. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 62–70, Osaka, Japan, 2016. COLING 2016 Organizing Committee.

[5] Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11:R86, 2010.

[6] Nancy Ide, James Pustejovsky, Eric Nyberg, Christopher Cieri, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. The Language

Application Grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

[7] Toru Ishida, Yohei Murakami, Donghui Lin, Takao Nakaguchi, and Masayuki Otani. Open Language Grid–Towards a Global Language Service Infrastructure. In *The Third ASE International Conference on Social Informatics*, Cambridge, Massachusetts, USA, 2014.

[8] Alexander Koplenig. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets: Reconstructing the composition of the German corpus in times of WWII. In *Digital Scholarship in the Humanities*, volume 32, 2016.

[9] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[10] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword Fifth Edition LDC2011T07, Linguistic Data Consortium, Philadelphia, 2011.

[11] Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLOS ONE*, 10(10):1–24, 10 2015.

[12] James Pustejovsky and Nancy Ide. Enhancing Access to Media Collections and Archives Using Computational Linguistic Tools. In *Corpora in the Digital Humanities (CDH)*, Bloomington, Indiana, 2017 (this volume).

[13] Erik Tjong Kim Sang. Finding Rising and Falling Words. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 2–9, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[14] Christopher Welty and Nancy Ide. Using the right tools: Enhancing retrieval from marked-up documents. *Computers and the Humanities*, 33(1-2):59–84, 1999.

[15] Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. Extracting Social Networks from Literary Text with Word Embedding Tools. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 18–25, Osaka, Japan, 2016. COLING 2016 Organizing Committee.