The Language Application Grid and Galaxy

Nancy Ide, Keith Suderman
Vassar College

James Pustejovsky, Marc Verhagen
Brandeis University

Christopher Cieri
Linguistic Data Consortium

Eric Nyberg
Carnegie-Mellon University
The LAPPS Grid

• A framework to
  – enable language service discovery, composition, and reuse
    • For both NLP researchers and others (who may use pre-developed composite services)
    – promote sustainability, manageability, usability, and interoperability of NLP components
• Based on the service-oriented architecture (SOA)
  – Web-oriented version of the “pipeline” architecture for sequencing loosely-coupled linguistic analyses
The LAPPS Grid

- Provides **interoperable access** to
  - Wide array of NLP processing tools and components
  - Language resources such as mono- and multi-lingual corpora and lexicons
- Enables pipelining tools to create **custom NLP applications and “black box” composite services**
- Provides an **open advancement (OA) framework** for component- and application-based evaluation
- Will enable **easy navigation through licensing issues**
- Actively pursuing creation of an **interoperable global network of grids and frameworks**
LAPPS Grid Overview

Galaxy workflow engine

LAPPS/Galaxy instance
User data

Directory of Tools, Resources, Usage Info

Data, Task, & Results Repository

User Server
Local Services
Local Data
Docker

Vassar Server
Service Grid Manager
Web Services

Brandeis Server
Service Grid Manager
Web Services

CMU Server
Evaluation/IAA Services

LDC Server
Data Delivery Services
Interoperability

• LAPPS Interchange Format (LIF)
  – allows services to exchange information
  – Syntactic interoperability
    • handled by JSON-LD
    • enforced by the LIF JSON schema
  – Semantic interoperability
    • enhanced by using the Linked Data aspect of JSON-LD to link to the LAPPS Web Services Exchange Vocabulary
      ➢ Not yet-another-repository! Linked to others where possible
**Definition**
A string of one or more characters that serves as an indivisible unit for the purposes of morpho-syntactic labeling (part of speech tagging).

**Similar to**
http://www.isocat.org/datcat/DC-1403

**URI**
http://vocab.lappsgrid.org/Token

**Metadata**

<table>
<thead>
<tr>
<th>Properties</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>posTagSet</td>
<td>String or URI</td>
<td>The definition of the tag set used by the part-of-speech tagger.</td>
</tr>
</tbody>
</table>

**Metadata from Annotation**

<table>
<thead>
<tr>
<th>Properties</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>producer</td>
<td>List of URI</td>
<td>The software that produced the annotations.</td>
</tr>
<tr>
<td>rules</td>
<td>List of URI</td>
<td>The documentation (if any) for the rules that were used to identify the annotations.</td>
</tr>
</tbody>
</table>

**Properties**

<table>
<thead>
<tr>
<th>Properties</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pos</td>
<td>String or URI</td>
<td>Part-of-speech tag associated with the token.</td>
</tr>
<tr>
<td>lemma</td>
<td>String or URI</td>
<td>The root (base) form associated with the token. URI may point to a lexicon entry.</td>
</tr>
<tr>
<td>tokenType</td>
<td>String or URI</td>
<td>Sub-type such as word, punctuation, abbreviation, number, symbol, etc. Ideally a URI referencing a pre-defined descriptor.</td>
</tr>
<tr>
<td>orth</td>
<td>String or URI</td>
<td>Orthographic properties of the token such as LowerCase, UpperCase, UpperInitial, etc. Ideally a URI referencing a pre-defined descriptor.</td>
</tr>
<tr>
<td>length</td>
<td>Integer</td>
<td>The length of the token</td>
</tr>
<tr>
<td>word</td>
<td>String</td>
<td>The surface string in the primary data covered by this Token.</td>
</tr>
</tbody>
</table>
Logical flow
(client-server communication not represented)

Data source

GATE service

GATE service

UIMA service

UIMA service

OpenNLP service

OpenNLP service

Stanford NLP service

Stanford NLP service

JSON-LD output

Converter to LIF JSON-LD

Converter from LIF JSON-LD

LAPPS services for OpenNLP and Stanford NLP tools are wrapped to produce and consume JSON-LD
All tools’ input/output formats mapped into and out of LIF

Linguistic categories etc. mapped to WSEV
LAPPS Web Service Exchange Vocabulary

• Specifies a terminology for a core of linguistic objects and features exchanged among NLP tools that consume and produce linguistically annotated data
• Linked wherever possible to existing repositories such as ISOCat (CLARIN Concept Repository), schema.org, FoLiA categories, etc.
• References in JSON-LD representation point to URIs providing definitions for specific linguistic categories in the WS-EV
Current collaborations/projects

- **Federation of Service Grids**
  - LAPPS Grid, Language Grid (Kyoto University, Japan), NECTEC (Thailand), University of Indonesia, Xinjiang University (China), ELRA Grid
  - Access to all tools, applications, and resources on any grid through any portal

- **LAPPS/CLARIN**
  - CLARIN/WebLicht (Tubingen) and LINDAT/CLARIN (Prague)
  - Mellon Foundation proposal to create a trust network between LAPPS and CLARIN

- **OpenMinted**
  - Advisory board—work together on harmonization

- **LAPPS Grid used in**
  - DARPA LORELEI project for under-resourced languages
  - HathiTrust Research Center (HTRC) text mining project
  - Multi-day course for government analysts
  - Undergraduate and graduate CL courses at CMU, Brandeis, Vassar
  - IBM Watson
LAPPS recently adopted the **GALAXY workflow engine** as a front end for construction of pipelines etc.

- open, web-based platform developed for computational genomics/biomedical research

http://galaxyproject.org
Why Galaxy?

- **Accessible**: Accommodates users with a broad range of expertise (non-computational to expert programmer)

- **Reproducible**: Galaxy captures information so that any user can understand and repeat a complete computational analysis

- **Transparent**: Users share and publish analyses via the web and create interactive, web-based documents that describe a complete analysis

- **Well-developed, supported, open!**
Galaxy

• Available as
  – A free public web server supported by the Galaxy Project (Johns Hopkins, Penn State)
    • Includes widely used bioinformatics tools
    • Users save histories, workflows, and datasets on the server, all can be shared with others
  – Open source software that can be downloaded and installed locally or in a cloud, and customized to address specific usages
  – Public web servers hosted by other organizations
Multiple options for running a LAPPS/Galaxy instance

- Use the LAPPS/Galaxy web interface
  - http://galaxy.lappsgrid.org
- Create a local Galaxy instance:
  - Clone our fork of the Galaxy project, or
  - Run the LAPPS Grid appliance
    - a series of docker images that is a self-contained vm running Galaxy and all LAPPS services
    - useful when privacy required, no network connection available, etc.
- Create an instance in the cloud
  - Useful for large datasets, computationally intense applications
Replicability and Sharing

• The field of NLP research and development has been plagued by a chronic lack of **replicability** of results
  – A great deal of re-inventing of the wheel and **wasted effort**
  – **Evaluation** of results hampered when details of a study (including versions and parameters for data, software) are not included in papers

• The field of NLP is still hampered by a lack of widespread **sharing** of resources that are the basis of research results
Galaxy as Promoter of Open, Replicable Research

• The field of NLP research and development has been plagued by a chronic lack of replicability of results and information about provenance

• Galaxy provides for:
  – automatic recording of inputs, tools, parameters and settings used for each step in an analysis in a publicly viewable history
  – sharing datasets, histories, and workflows via web links
  – creation of custom web-based documents to communicate about an experiment or result

• Encourages open publication, sharing of data and results

• Fosters replicability and reuse by providing a physical infrastructure to enable it
LAPPS/Galaxy

• LAPPS/Galaxy components are LAPPS web services
• Access to 100+ LAPPS services plus those of federated partners
• Interoperability
  – LAPPS services among each other
  – LAPPS services and Galaxy components
    • handled by LAPPS converters
**LAPPS Grid**

A Framework for Rapid Adoption and Reuse.

**Work In Progress**

Many of the services here are undergoing active development and their behaviour is likely to change without notice.

Welcome to the LAPPS Grid Galaxy instance. Through this Galaxy instance you can:

1. Fetch documents from the MASC or Gigaword corpora.
2. Create processing pipelines with tools from:
   1. GATE
   2. Apache OpenNLP
   3. Stanford NLP

**Simple Tutorial**

If you have a good understanding of how Galaxy works you can run the following tools in order:

1. Get data -> MASC
2. From the GATE menu ->
   1. Tokenizer
   2. Sentence Splitter
   3. Part of speech tagger
3. From the History panel select ->
   1. Edit attributes
   2. Convert Format (there is only one converter, so just run it)
4. Tools -> Word Count
5. Expand the output select the Visualization icon and then Ch plots
Workflow construction

LAPPS provides interoperability among...

... GATE tools

... Stanford tools

... OpenNLP tools

... others!
Step-by-step analysis

Download data directly from web or upload files from your disk

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

- **Name**: New File
- **Size**: 0.1 KB
- **Type**: Auto-detect
- **Genome**: unspecified
- **Settings**: 

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

7. Click the draw button to marvel at the simple word count histogram ;-)
1 job has been successfully added to the queue - resulting in the following datasets:

5: Stanford Dependency Parser v2.0.0 on data 4

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
Online Visualization of LappsGrid

LappsGrid, Version 0.3.0, May 2015

Brat Display

Tool Output

```
1 {  
    "discriminator": "http://vocab.lappsgrid.org/ns/media/jsonld",
    "payload": {
      "@context": "http://vocab.lappsgrid.org/context-1.0.0.jsonld",
      "metadata": {},
      "text": {
        "#value": "Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state."
      },
      "views": []
    }
  }
```

1 job has been successfully added to the queue - resulting in the following datasets:

6: Stanford Parser v2.0.0 on data 5

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.

```
1. Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.
2. (ROOT [149.288])
3. (SINV [146.125])
4. (VP [28.611] (VBG Binding))
5. (PP [16.520] (TO to))
6. (NP [14.049] (NNP causes))
7. (NP [104.025])
8. (NP [22.775] (DT a) (JJ conformational) (NN change))
9. (PP [78.902] (IN of))
10. (NP [77.375])
11. (NP [28.141] (DT the) (NN ras) (NN protein))
12. (PP [58.293] (IN into))
13. (WHNP [1.447] (WDT that))
14. (S [47.386])
15. (VP [47.110] (VBZ puts))
16. (NP [15.984] (NNP Ras))
17. (PP [20.534] (IN into))
18. (NP [18.071] (DT the) (JJ active) (NN state)))))))
```
Evaluation in LAPPS/Galaxy

- CMU has implemented services for state-of-the-art Open Advancement techniques.
- Enables rapid identification of
  - frequent error categories within modules and documents
  - which module(s) and error type(s) have the greatest impact on overall performance
- Used in the development of IBM’s Watson to achieve steady performance gains over the four years of its development.
Open Advancement in a Nutshell

• Analyzes results in/from alternative pipelines

- Can be comparison to gold standard, or comparison to another pipeline or pipelines
- Potentially any number of pipelines can be compared
  - CMU working on methods for finding an optimal solution among all multiple possible paths
Parallel workflows
### MASC gold standard vs. Stanford NEs

**Precision:** 0.391  
**Recall:** 0.473  
**F1:** 0.428

<table>
<thead>
<tr>
<th>Reference Outputs</th>
<th>Predicted Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Start</strong></td>
<td><strong>End</strong></td>
</tr>
<tr>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>99</td>
</tr>
<tr>
<td>104</td>
<td>110</td>
</tr>
<tr>
<td>145</td>
<td>151</td>
</tr>
<tr>
<td>389</td>
<td>402</td>
</tr>
<tr>
<td>429</td>
<td>438</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>490</td>
<td>494</td>
</tr>
<tr>
<td>647</td>
<td>657</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>738</td>
<td>743</td>
</tr>
<tr>
<td>899</td>
<td>904</td>
</tr>
</tbody>
</table>
Potential benefits of LAPPS/Galaxy collaboration

• Galaxy contains a huge number of tools for analyzing genomic and other biomedical data
• LAPPS includes tools to perform NLP analyses
• Combining LAPPS services with Galaxy tools can allow for analysis of data mined from the vast stores of biomedical literature (Biomed, PubMed, PLOS, etc.)

• BIONLP meets bio-analysis!
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state. GTP-bound ras binds to the raf protein kinase. This binding of raf to ras has the effect of activating the raf kinase and localizing the raf kinase to the cell membrane. Activated raf now phosphorylates and activates the Mek1 kinase. The Mek1 kinase then phosphorylates the ERK kinase on both threonine and tyrosine residues which activate ERK kinase activity. The phosphorylated ERK protein then translocates to the nucleus where it regulates gene expression in part by phosphorylating the Elk1 transcription factor. Phospho-Elk then upregulates the gene expression of target genes such as the proto-oncogene c-fos. The entire signaling cascade is terminated by the intrinsic GTPase activity of ras which hydrolyzes the bound GTP into GTP, thus returning ras to the GDP bound state where it releases bound raf. The GTPase activity of ras is accelerated by interaction with another protein called GAP. The oncogenic rasv12 mutant has diminished GTPase activity and therefore stays in the active GTP bound state constitutively. Deletion of GAP or the related NF1 genes will also enhance ras activity by slowing the rate of ras-GTP hydrolysis.
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.
Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.
GTP-bound ras binds to the raf protein kinase.

Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.
GTP-bound ras binds to the raf protein kinase.

Binding to GTP causes a conformational change of the ras protein that puts Ras into the active state.

GTP-bound ras binds to the raf protein kinase.
LAPPS/GALAXY Demo

P31
LR Infrastructures and Architectures
Thursday May 26
14:55-16:35
Poster Area 1
Thank You