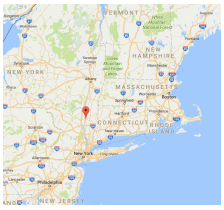
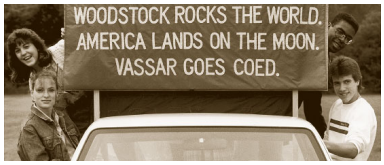
Mining Biomedical Publications with the Language Applications Grid

Nancy Ide • Department of Computer Science • Vassar College





Vassar College est. 1865
Poughkeepsie, NY USA

Need for text mining

- Number of new scientific publications is growing rapidly
- New terms (genes, proteins, chemical compounds, drugs) are constantly created
- Impossible to manage such an information overload!

Scientific Information Overload

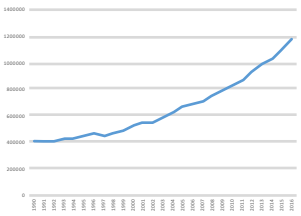
The global research community generates ~2.5 million new scholarly papers per year (English only)

A new research paper is published every 12 seconds

70,000 papers published on a single protein

Challenge to scientists	Keep updated on new developments, paper writing, project proposal preparation, paper reviewing, peer assessment, etc.

Publications Added to PubMed 1990-2016



Total for all years close to 30 million articles

PubMed is now accumulating over 1,000,000 new entries every year

Need for text mining

- Increased availability of full text – Information retrieval is an insufficient solution
- Bio-databases, controlled vocabularies and bio-ontologies encode only small fraction of information
- Most information is in textual form – unstructured data
- Automated aids are needed!

The Solution: NLP!

- Techniques that analyze, understand and generate language (free text, speech)
- Multidisciplinary field: information technology, computational linguistics, AI, statistics, psychology, language studies, etc.
- Create computational models of language
- Strongly language-dependent
- Explore grammatical, morphological, syntactical and semantic features of well-structured language
- Methods: Statistical analysis, machine learning
- Statistical analysis of these features in large text collections is the basic approach used in NLP

Some NLP Tasks Relevant for Exploring Biomedical Texts

1. **Information Retrieval** yields all relevant texts
   - Gathers, selects, filters documents that may prove useful
   - Finds what is known
2. **Information Extraction** extracts facts and events of interest to user
   - Finds relevant concepts, facts about concepts
   - Finds only what we are looking for
3. **Data Mining** discovers unsuspected associations
   - Combines and links facts and events
   - Discovers *new* knowledge, finds new associations

NLP Applications

- Involve a sequence of processing steps
  - Provide increasingly sophisticated layers of analysis
  - Exact sequence varies for different tasks

Relation Extraction
Co-reference resolution
Named Entity Recognition
Parsing/ Chunking
POS-tagging
Sentence splitting
Tokenization
Raw text

Basic Processing Steps

- Most applications require at least the following:

POS-tagging	my_PRP dog_NN 's_POS fleas_NNS are_VBP n't_RB gone_VBN ._.
Tokenization	my dog 's fleas are n't gone .
Sentence splitting (n/a)	My dog's fleas aren't gone.
Raw text	My dog's fleas aren't gone.

NLP Applications

- Some are an "all-in-one" black box
  - E.g., Noa's software for negation identification
  - Modules for tokenization, sentence splitting, POS-tagging, maybe other things are inside, invisible to user
- But it is often desirable to compose your own application from independent modules
  - Create a new application
  - Customize certain modules
  - Test the effect of different modules on the quality of results
  - Experiment with different resources (e.g. lexicons used by NER module)

NLP Pipelines (Workflows)

- Chains of NLP tools to perform a given task

- A number of NLP *frameworks* provide environments for chaining tools together into pipelines
  - E.g. General Architecture for Text Extraction (GATE) and Unstructured Information Management and Analysis (UIMA)
- Others provide sets of modules designed to work together (API) to be called from a user program
  - Stanford NLP tools, OpenNLP, Natural Language Toolkit (NLTK), LingPipe, DKPro…

Typical Scenario

- A scientist wants to apply text mining techniques to find articles including references to certain entities (e.g., proteins, genes) and their interactions
  - Knows nothing about NLP or Computer Science
  - Unfamiliar with NLP technologies
- Searches for NLP software that might help

Typical Scenario

- Finds existing tools and frameworks that are freely available

Not to mention several commercial (i.e., pricey) options

- Questions
  - Do these things all do the same thing, or do they differ in some way?
  - Do some work better than others?
  - Are some easier to use than others?
  - How does one choose?

Confused scientist

Issues

- Most of these tools provide general NLP support, not geared to BioNLP
- Why does this matter?

Characteristics of biomedical literature

1

Heavy use of domain specific terminology (12% biochemistry related technical terms)

- Examples: chemoattractant, fibroblasts, angiogenesis

2

Constant introduction of new terms and short forms or abbreviations

Characteristics of biomedical literature

3

Polysemic words

Examples: APC: (1) Argon Plasma Coagulation (2) Activated Protein C; or teashirt: (1) a type of cloth (2) a gene name (tsh).

4

Heavy use of acronyms

Examples: Activated protein C (APC) , or vascular endothelial growth factor (VEGF)

Characteristics of biomedical literature

5a

Term Ambiguity

Gene terms may be also common English words

- *BAD* human gene encoding BCL-2 family of proteins (*bad news, bad prediction)*

5b

Term Ambiguity

Gene names are often used to denote gene products (proteins)

- *suppressor of sable* is used ambiguously to refer to either genes and proteins

Characteristics of biomedical literature

6

Typographical variants (e.g. in writing gene names)

- Example: TNF-alpha and TNF alpha (without hyphen)

7

Most words have low frequency (data sparseness)

No Problem!

Lists of NLP Tools for Biomedical texts

- http://bionlp.org
- http://biocreative.sourceforge.net/bionlp_tools_links.html
- http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7784950
- http://www.nactem.ac.uk/index.php (but also lots of other things)

However…

- Many of these tools are difficult to install, configure, and use without some computational expertise
- Even more difficult to modify or adapt without computational expertise and some knowledge of NLP
- Also: which tools performing the same task perform best and/or are best suited to a given task?

Another Sneaky Underlying Problem

- Input and output of tools from different sources differ dramatically!!!
- Often demands significant effort and expertise to adapt tools from different sources to work together
  - …if it is possible at all
- I.e., tools are not interoperable

Lack of Interoperability

- Software waste cycle
  - Because of the difficulties of adapting software developed at other sites to one's data and/or purpose, researchers often write their own versions
  - Huge waste of time and effort to reinvent the wheel
- Increasing availability of open tools and data
  - Open tools and data developed at a single site (e.g. Stanford tools) usable together with no modification
  - BUT
    - Want to be able to use tools from different sites together with minimal or no effort to adapt
      - Important for Open Advancement approach to development
    - Want to be able to process data in various formats with minimal or no effort to adapt

What is Needed

1. Develop/ provide access to a range of freely available advanced text mining tools specially tailored to scientific publications
2. Enable scientists to easily use these tools without having to be a computer scientist or an expert in NLP

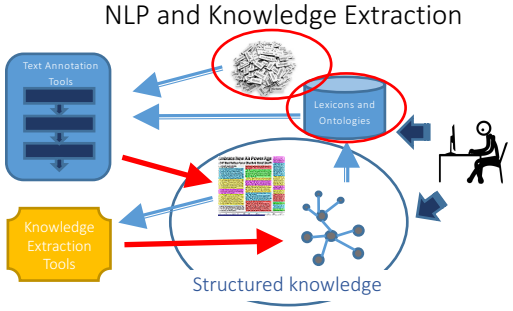
   Interoperability is key!
3. Enable scientist to easily adapt existing solutions to specific domains or problems without having to be a computer scientist or an expert in NLP

Tools are Not Enough

BioNLP also needs language resources:

- Large bodies of scientific publications that can be searched and mined for information and knowledge
- Large bodies of annotated scientific publications that can be used to develop language models (e.g. via machine learning)
- Lexicons and Ontologies of biomedical terms to assist in entity recognition etc.

NLP and Knowledge Extraction



Biomedical Literature Repositories

- Electronically accessible to the public through the web
  - Centralized institutional (for example, PubMed) or academic (for instance, Highwire Press and Hollis) repositories of peer-reviewed articles or article abstracts.
  - Article collection repositories hosted by publishers (for example, BioMed Central and EMBASE).
  - Online access to indexed scholar articles retrieved through web spiders and crawlers (for example, Google Scholar and Scirus).
- Masic and Milinovic (2102) provide one overview of world literature databases for the Biomedical field

Despite all this, a corpus resource bottleneck

1

Lack of freely available large-scale, richly annotated corpora to support

- Training of ML algorithms
- Development of computational grammars
- Evaluation of text mining components

2

Existing corpora in many different formats

- Often necessary to extract text from PDF (not always accurate!)
- Formats such as various home-grown XML, JSON, "BIO", …

3

Annotations may be named and/or structured very differently

- Annotations for the same phenomenon have to be transduced to be used together
- Sometimes there is no direct mapping
- Not interoperable!

Knowledge Resources for BioNLP

1

Lexical and terminological resources

- Lists of terms / lexical entries

2

Ontological resources

- Concepts with definitions, sometimes includeing synonyms
- Relations between the concepts, typically is-a and has-part

Table 1. Some knowledge sources for biomedical natural language processing.

Informatics for Integrating Biology and the Bedside (i2b2 - https://www.i2b2.org/)	National Center for Biomedical Computing with focus on translational research that facilitates and proves data sets for clinical natural language processing research
Gene Ontology (https://www.geneontology.org)	Controlled vocabulary with relationships including partonymy and inheritance, designed for describing gene functions, broadly construed
Entrez Gene (https://www.ncbi.nlm.nih.gov/gene)	Source for gene names, symbols, and synonyms; also the source for GeneRIFs and SUMMARY fields
PubMed/MEDLINE (https://www.ncbi.nlm.nih.gov/pubmed)	The National Library of Medicine's database of abstracts of biomedical publications (MEDLINE) and search interface for accessing them (PubMed)
Unified Medical Language System (https://www.nlm.nih.gov/research/umls/)	Large lexical and conceptual resource, including the UMLS Metathesaurus, which aggregates a large number of biomedical and some genomic vocabularies
SWISSPROT (https://www.uniprot.org/)	Database of information about proteins with literature references, useful as a gold standard
PharmGKB (https://www.pharmgkb.org/)	Database of relationships between a number of clinical, genomic, and other entities with literature references, useful as a gold standard
Comparative Toxicogenomics Database (https://ctdbase.org/)	Database of relationships between genes, diseases, and chemicals, with literature references, useful as a gold standard

Various terminological resources, data sources, and gold-standard databases for biomedical natural language processing.
doi:10.1371/journal.pcbi.1003044.t001

Cohen KB, Hunter LE (2013) Chapter 16: Text Mining for Translational Bioinformatics. PLOS Computational Biology 9(4): e1003044. https://doi.org/10.1371/journal.pcbi.1003044
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003044

Are the existing knowledge resources sufficient for text mining?

NO!

- Limited lexical and terminological coverage of biological sub-domains
- Terminological variation and complexity of names
  - Variation occurs in controlled vocabularies and texts but discrepancy between the two
  - Exact match methods fail to associate term occurrences in texts with databases

Furthermore…

- The same interoperability problem exists for resources!
  - Different physical formats
    - PDF, XML, plain text…
    - Extraction of text from PDF can be unreliable
  - Different representations for annotations
    - Different physical formats
      - XML, JSON, brackets, BIO

BIONLP2016 Shared Task Data Sets

```
(ncmod _ structures_2 Chromosomal_1)
(ncmod _ testosteroni_5 Pseudomonas_4)
(dobj of_3 testosteroni_5)
(ncmod _ structures_2 of_3)
<c> | | LRB | LRB Chromosomal | JJ | N/N structures | NNS | N
of | IN | (NP\NP)/NP Pseudomonas | NN | N/N
testosteroni | NN | N . | . |
<c> | | NN | N . | . |.
```

```
CDI     199 202   ORGANIZATION
CDI     375 378   ORGANIZATION
CDI     426 429   ORGANIZATION
CDI     583 586   ORGANIZATION
CDI     713 716   ORGANIZATION
CDI     864 867   ORGANIZATION
Olmsted County    1001 1015 LOCATION
Minnesota 1017 1026 LOCATION
```

```
<TERM_EXTRACTION_RESULTS>  <LIST_TERM_CANDIDATES>
<TERM_CANDIDATE MNP_STATUS="1">
<ID>term665</ID>
<FORM>control</FORM>       <LEMMA>control</LEMMA>
<MORPHOSYNTACTIC_FEATURES>
<SYNTACTIC_CATEGORY>NN</SYNTACTIC_CATEGORY>
</MORPHOSYNTACTIC_FEATURES>
<HEAD>term665</HEAD>
<NUMBER_OCCURRENCES>8</NUMBER_OCCURRENCES>
<LIST_OCCURRENCES>
<OCCURRENCE>
<ID>occ4746</ID>
<MNP>B</MNP>
<DOC>25</DOC>
<SENTENCE>3</SENTENCE>
<START_POSITION>160</START_POSITION>
<END_POSITION>167</END_POSITION>
</OCCURRENCE>
```

```
The         The         DT      B-NP    O
etiologic   etiologic   JJ      I-NP    O
and         and         CC      I-NP    O
epidemiologic           epidemiologic   JJ      I-NP    O
spectrum    spectrum    NN      I-NP    O
of          of          IN      B-PP    O
bronchiolitis           bronchiolitis   NN      B-NP    O
```

You see the problem.

Added to the other problems, this makes NLP very difficult for scientists to exploit.

Portions of the preceding slides based on:

Krallinger, Martin, Valencia, Alfonso, and Hirschman, Lynette. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology* 9:2, 2008.

Tutorial: Text Mining for Biomedicine: Techniques & tools. Slides by Sophia Ananiadou, National Centre for Text Mining

Biomedical Natural Language Processing, by Kevin Bretonnel Cohen, Dina Demner-Fushman, Benjamins Publishing.

Nancy Ide, Keith Suderman
Vassar College

James Pustejovsky, Marc Verhagen
Brandeis University

Christopher Cieri, Denise DiPersio, Jonathan Wright
Linguistic Data Consortium (Penn)

Eric Nyberg, Di Wang
Carnegie Mellon University

The Language Applications (LAPPS) Grid

What is the LAPPS Grid?

- US National Science Foundation-funded project
- Collaborative among Vassar College, Brandeis University, University of Pennsylvania, and Carnegie Mellon University
- Goal: Provide an infrastructure that facilitates
  - Retrieving large text collections from providers (Linguistic Data Consortium) and other holders and repositories
  - Devising pipelines (workflows) of **interoperable** web services that automatically annotate the data, provide evaluation metrics for the results, etc.
  - Saving, storing, and sharing pipelines and results for later use by yourself or others

More Goals

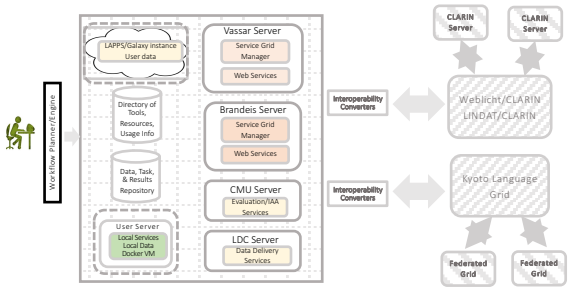
- Provide access to a broad range of NLP tools that are interoperable
- Enable creation of custom applications and "black box" composite services
- Provide an open advancement (OA) framework for component- and application-based evaluation
- Enable easy navigation through licensing issues
- Actively pursue creation of an interoperable global network of grids and frameworks

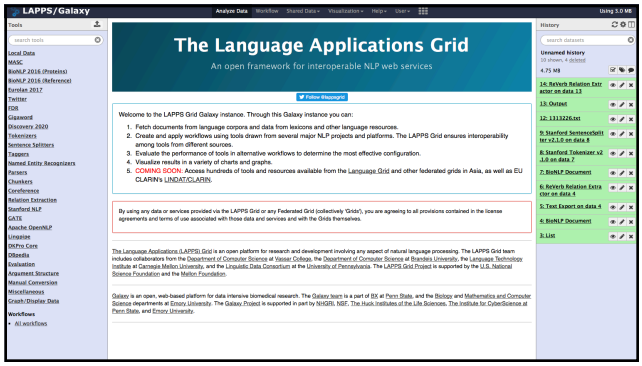
LAPPS/Galaxy Interface

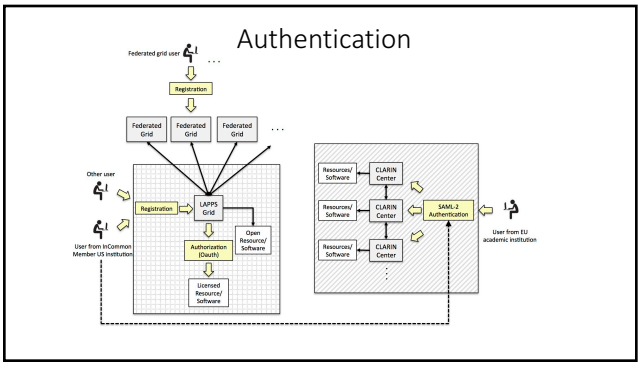
Galaxy
http://galaxyproject.org

- Galaxy is an open, web-based platform designed primarily for computational biomedical research
  - **Accessible:** Users without programming experience can easily specify parameters and run tools and workflows
  - **Reproducible:** Galaxy captures information so that any user can repeat and understand a complete computational analysis
  - **Transparent:** Users share and publish analyses via the web and create interactive, web-based documents that describe a complete analysis
- The LAPPS Grid uses the GALAXY framework as a vehicle to combine services of the Language Application Grid
  - Text processing pipelines, components wrapped as services, visualization of component output, evaluation of alternate pipelines, saving and sharing workflows, etc.

LAPPS Grid Overview







LAPPS/GALAXY

Multiple options for running the LAPPS/Galaxy instance:

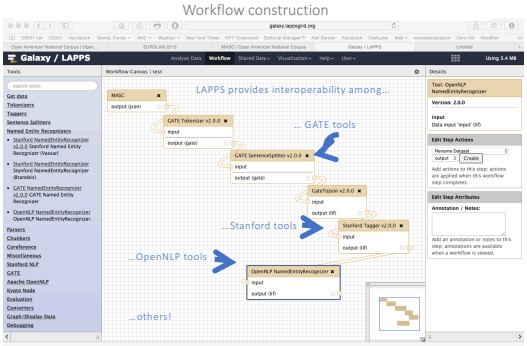
1. Use the LAPPS/Galaxy web interface
   - http://galaxy.lappsgrid.org
2. Create a local Galaxy instance including:
   - All of Galaxy, or
   - The Galaxy "NLP Flavor" with only LAPPS tools
3. Create a docker image that is a self-contained vm running LAPPS/Galaxy
   - Useful when privacy required, no network connection available, etc.
4. Create a Galaxy instance in the cloud
   - Useful for large datasets, computationally intense applications
   - Useful for tutorials!
   - **HERE: https://jetstream.lappsgrid.org**

Interoperability

- The Galaxy bio tools are not interoperable
  - Multiple converters among genomic formats where it is possible to provide them
- LAPPS Grid services are interoperable!

Format Converters

Workflow construction



Brief digression: What is interoperability?

INTEROPERABILITY DEFINED

in·ter·op·er·a·bil·i·ty

\ˌin-tər-ˌä-p(ə-)rə-ˈbi-lə-tē\

noun

ability of a system (as a computer system or software) to exchange and make use of information.

INTEROPERABILITY DEFINED

- "**Interoperability** is a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, in either implementation or access, without any restrictions."

Wikipedia

More precisely…

- Data is in a format that is immediately readable and processable by a given piece of software
- Output of a given piece of software can be read into and operated upon by another piece of software without modification

Two levels of interoperability

Syntactic interoperability

- Relies on specified data formats, communication protocols, and the like to ensure communication and data exchange

Semantic interoperability

- The content of the information exchange requests are unambiguously defined: what is sent is the same as what is understood

Syntactic interoperability

- Systems involved can process the exchanged information, but no guarantee that the interpretation is the same
  - EXAMPLE: Data in XML format is readable by software designed to recognize XML

Software recognizes <..> as a tag…

…but it has no idea what "NAME" means–i.e., how that tag's content may need to be processed differently from <P> or <DOC>

Semantic interoperability

- Two systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results via deference to a common information exchange reference model

Software knows the "meaning" of NAME–i.e., can perform specific steps to process it differently from P or DOC

Interoperability for language data

- Syntactic interoperability achieved with common physical formats
  - Many options: One sentence per line, part-of-speech tag appended to word, XML…
- Semantic interoperability achieved with common definitions for labeled data
  - E.g., labels like *noun*, *person, date* mean the same to both systems

Not easy! There can be subtle differences of opinion (e.g., should "in the future" be labeled as a DATE? Is "the White House" a LOCATION or an ORGANIZATION in a phrase like "The White House said today…")?
Let alone that people do not agree on the exact definition of noun…

Obstacles

- Difficult to identify a single representation format that accommodates all kinds of language data and annotations
- Difficult to get the community to agree, adopt a single standard
- Need to accommodate legacy data and tools using other formats

Current solution

- 30 years of development have led to reasonable convergence of practice
- Key idea:
  - Instead of defining a single solution, design a universal "pivot" into and out of which other schemes can be easily mapped
  - For physical formats, requires that the pivot is a serialization of a common abstract data model (directed acyclic graph)
    - This model underlies UML, ER diagrams, RDF, JSON and JSON-LD, XML, semantic and other kinds of networks…
  - For semantics, provide a common structured set of terms to which other schemes can be mapped

Once again, this is non-trivial

So, How Does the LAPPS Grid Enable Interoperability?

- LAPPS Interchange Format (LIF)
  - Format that allows web services to exchange detailed information about data and its annotations
  - 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

Web Service Communication in LAPPS

1

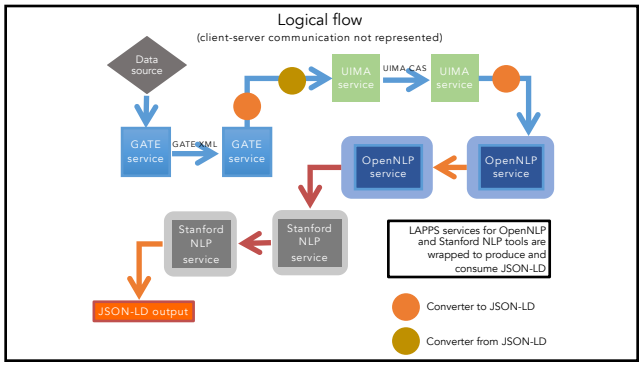
Each service in the LAPPS Grid publishes metadata:

- a **discriminator** (type) : tells how to interpret the payload
- a **payload** (typically a utf-8 string)

2

LAPPS uses JSON-LD as its standard format for the payload

- **Converters** to and from JSON-LD for services that deliver in other formats
- Some LAPPS services are wrapped to **produce and consume JSON-LD**



LAPPS Grid Web Service Exchange Vocabulary

- No accepted standard for module description or input/output interchange in the language application domain
- LAPPS Web Service Exchange Vocabulary (WS-EV)
  - specifies a terminology for a core of linguistic objects and features exchanged among NLP tools that consume and produce linguistically annotated data
  - addresses a need within the community to identify a standard terminology and indicate the relations among them

Design Principles

01 Orthogonal design
- Only one entry per concept

02 Lightweight
- Easy to find on the web and reference

03 Flexible
- Use what you need, add what you need

04 (Arbitrary) decisions about what goes where
- Map to this for exchange only
- Not confined to the WS-EV terminology or organization internally

User definition of objects and features

	Many web services will require definition of objects and properties not included in the WS-EV or elsewhere
Options	Provide a URI where a new term or other documentation is defined Add a definition to the WS-EV • Service providers use the name space automatically assigned to them at the time of registration • Avoids name clashes • Makes a distinction between general categories used across services and more idiosyncratic categories

Implementation

- **Bottom-up approach**
  - Define objects and properties as needed to accommodate LAPPS services as they are added to the LAPPS Grid
  - Avoids *a priori* development of a comprehensive standard linguistic type system
  - "Minimalist" strategy to provide a simple core set of objects and features
  - User capacity to add/replace objects and properties to allow for dynamic typing

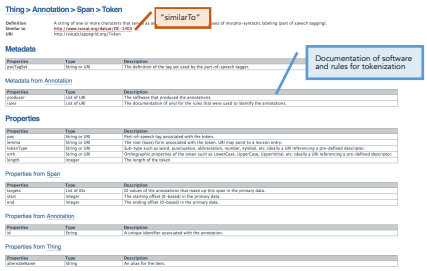
LAPPS WS-EV Repository

- http://vocab.lappsgrid.org
- Shallow hierarchy of elements
  - Inheritance

LAPPS Exchange Vocabulary Type Hierarchy

Thing: alternateName
- Annotation: id
  - Region: targets, start, end
    - Paragraph
    - Sentence: sentenceType
    - NounChunk
    - VerbChunk: vcType, tense, voice, neg
    - NamedEntity: category, type, gender
      - ~~Date: dateType~~
      - ~~Location: locType~~
      - ~~Organization: orgType~~
      - ~~Person: gender~~
    - Token: pos, lemma, tokenType, orth, length, word
    - Markable
  - Relation: label
    - GenericRelation: relation, arguments
    - SemanticRole: head, argument
    - Constituent: parent, children
    - Dependency: governor, dependent
  - Coreference: mentions, representative
  - PhraseStructure: constituents, root
  - DependencyStructure: dependencyType, dependencies
- Document: id, source, sourceType, encoding,
- language
  - TextDocument
  - AudioDocument

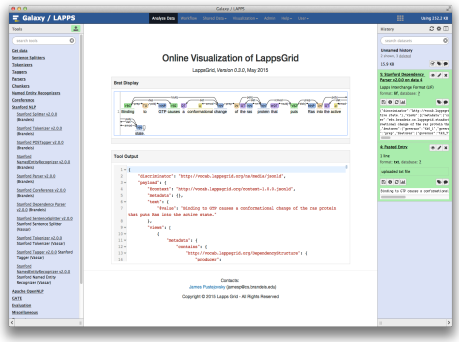
Spec for Token

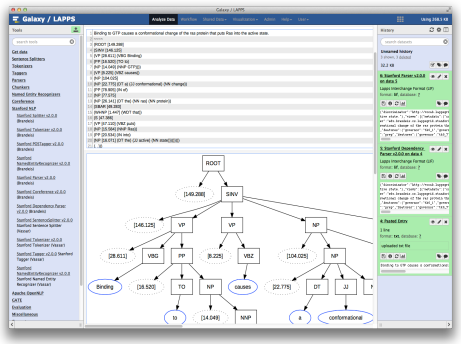


JSON-LD and WS-EV

- References in JSON-LD representation point to URIs providing **definitions** for specific linguistic categories in the WS-EV
- Also point to **documentation** for processing software and rules for processes such as tokenization, entity recognition, etc.
  - Often left unspecified in annotated resources
  - Not required for web service exchange in the LAPPS Grid
  - **BUT** inclusion of such references can contribute to better replication and evaluation of results in the field
  - **Promote best practice!**

JSON-LD and the LAPPS Exchange Vocabulary





EVALUATION IN LAPPS

- CMU has implemented services for state-of-the-art Open Advancement techniques
- **Provides an unprecedented tool for NLP development**
  - Could take the field to a new level of productivity
- 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

01

Evaluate multiple possible solutions (tool configurations) for a given problem

- Determine the optimal solution available using given components, resources, and evaluation data

02

Output of the optimal solution subjected to error analysis

- Identify the most frequent errors with the highest impact
- Consider possible enhancements
  - Aim to achieve the largest possible reduction in error rate by addressing the most frequent error types

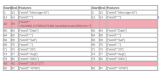
03

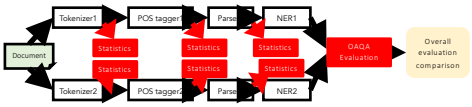
Evaluate performance of new configurations

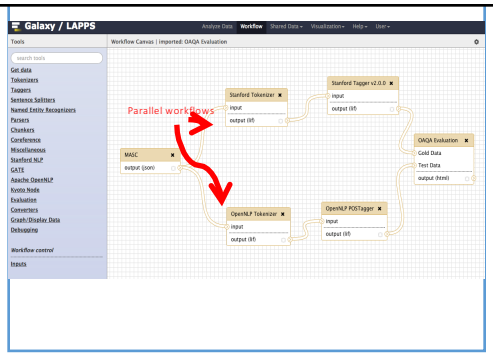
- Determine if a significant improvement has been achieved in comparison with prior baselines

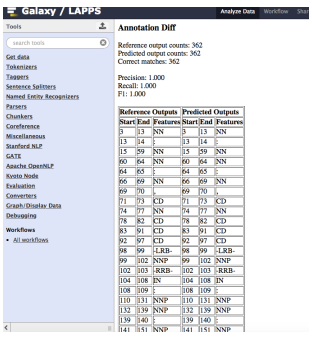
CMU OAQA

- Open advancement for Question Answering
- Analyzes results in/from alternative pipelines









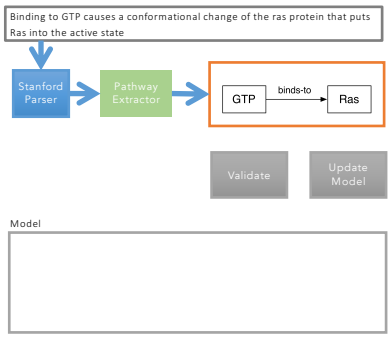
Reference Outputs			Predicted Outputs		
Start	End	Features	Start	End	Features
3	13	NN	3	13	NN
13	14	:	13	14	:
15	59	NN	15	59	NN
60	64	NN	60	64	NN
64	65	:	64	65	:
66	69	NN	66	69	NN
69	70	.	69	70	.
71	73	CD	71	73	CD
74	77	NN	74	77	NN
78	82	CD	78	82	CD
83	91	CD	83	91	CD
92	97	CD	92	97	CD
98	99	-LRB-	98	99	-LRB-
99	102	NNP	99	102	NNP
102	103	-RRB-	102	103	-RRB-
104	108	IN	104	108	IN
108	109	:	108	109	:
110	131	NNP	110	131	NNP
132	139	NNP	132	139	NNP
139	140	.	139	140	.
141	151	NNP	141	151	NNP

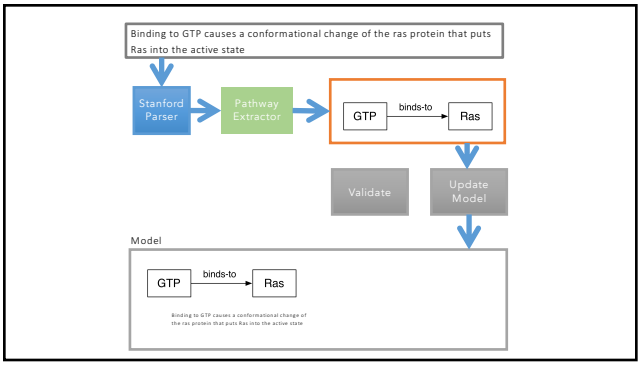
BENEFITS OF THE LAPPS/GALAXY COLLABORATION

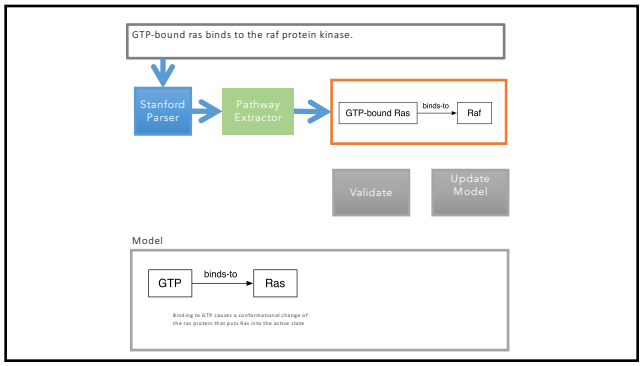
- Galaxy contains a huge number of tools for analyzing genomic and other biomedical data
- LAPPS includes tools to perform NLP analyses
- Combining LAPPS services with Galaxy tools can provide data mined from the vast stores of biomedical literature (Biomed, PubMed, PLOS, etc.)
  - interaction between model and observations (pathway steps), can be mined from new text
- BIONLP meets bio-analysis!

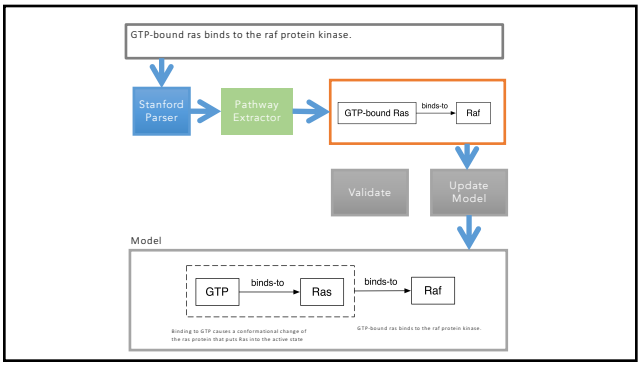
Example

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.









Current Activities

- New NSF grant
  - Collaboration between Vassar College and Penn State University (and other Galaxy Principal Investigators) to
    - Develop a robust and interoperable set of tools, ready-made workflows, etc. for mining biomedical publications
      - Rely on researchers to test and provide feedback as we develop
    - Provide seamless integration of text mining capabilities and the vast array of tools provided in Galaxy
- Collaboration with the US government Centers for Disease Control and Food and Drug Administration to adapt the LAPPS Grid for summarization and mining of clinical reports

Penn BioTokenizer

Biomedical NER
- Annotates proteins, DNA, RNA, cellLines, cellTYpes

Gene annotator

BioNLP-oriented Tools in the LAPPS Grid

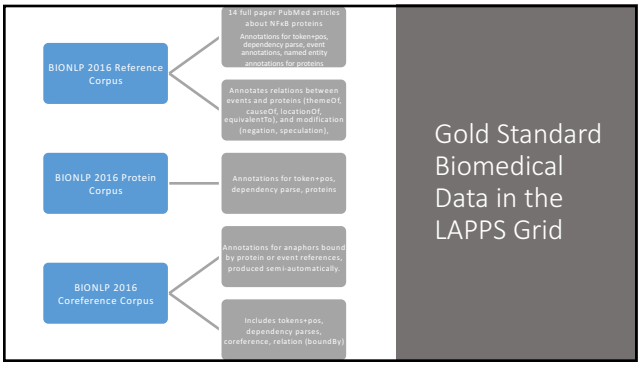
TimeML Events

LingPipe Dictionary-based NER

Several different NER modules, tokenizers, parsers, chunkers, etc.

Evaluation tools (Open Advancement)

Other LAPPS Grid Tools Useful for BioNLP

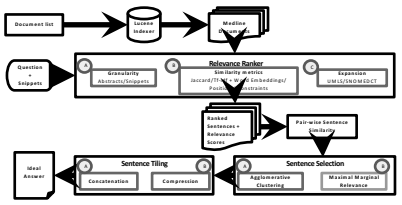


BioASQ

- Large-scale biomedical semantic indexing and question answering challenge, started in 2013
- Phase B of Task 5b of the BioASQ challenge: Biomedical Q/A
  - Involved answering the following types of questions
    - Factoid
    - List
    - Yes/ No
    - Summary
  - Exact answers for Factoid, List, Yes/ No
  - Ideal answers for all questions types
  - Five test sets, comprising of 100 questions each

Pipeline Architecture of Winning System

Carnegie Mellon University
Language Technologies Institute



Hands-on Exercises

01

Using the LAPPS Grid

- Learn how to import data, execute services, create workflows, visualize results, etc.
- Explore various tools, including parsers, named entity recognizers

02

Exploring biomedical data and tools

- Import data without annotations, annotate with LAPPS Grid tools
  - Tokenizer, protein annotator, event annotator
  - Dictionary-based NER

03

Evaluating automatically-produced results

- Use LAPPS Grid evaluation services to compare gold standard annotations with automatically-generated ones