SciKnowMine 2013 Workshop - Bridging BioNLP and Biocuration

PART I - SciKnowMine Overview

8:45 - 9:00 : Introductory remarks (Gully APC Burns)

9:00 - 9:30 : The SciKnowMine Project - Bridging BioNLP and Biocuration (Gully APC Burns)

9:30 - 10:00 : Mouse Genome Informatics (MGI): the challenge to develop NLP strategies and tools to support publication-based curation (Janan Eppig)

10:00 - 10:30 : The Colorado Richly Annotated Full Text corpus: A corpus linguist's perspective (Kevin B Cohen)

10:30 - 11:00 Break

PART 2 - Biomedical Databases : at the cutting edge of knowledge

11:00 - 11:30 : Reactome – Linking pathways, networks and disease (Robin Haw).

11:30 - 12:00 : OntoPub: An Ontology-Driven, Concept-Based Biomedical Literature Search Engine (Dongui Lui)

12:00 - 12:30 : Wiki-based community annotation (Jim Hu)

12:30 - 1:30pm : Lunch

PART 3 - BioNLP : computing over the complexity

1:30 - 2:00 : An NLP Voyage: Explorations with Information Extraction for Biocuration (Ellen Riloff)

2:00 - 2:30 : Mining Fulltext (Maximilian Haeussler)

2:30 - 3:00 : Clarifying ClearTK: A Case Study in Entity Tagging (James Gung)

3pm - 3:30pm : Break

PART 4 - Technological approaches applied to publishing space

3:30 - 4:00 : Mendeley's vision of biocuration (William Gunn)

4:00 - 4:30 : The BioCreative Evaluation initiative (Cecilia Arighi)

PART 5 - Forging the future - panel discussion

4:30 - 5:30 : Building Bridges: Moderated community discussion.

PART 6 - Evening Dinner

7:00 - 7:10 : Concluding remarks and wrap up

The SciKnowMine Project - Bridging BioNLP and Biocuration

Gully APC Burns, Intelligent Systems Division, Information Sciences Institute.

For biologists, text mining has always promised to alleviate the massive information overload endured by most working scientists who struggle to keep up with the scientific literature. In practice, even apparently simple text mining problems tend to contain challenging computer science problems of interest to experts in the field of Natural Language Processing (NLP). Although the two communities of researchers could greatly benefit from working with each other, each have their own set of goals which do not align well with their counterparts. The motivation for this workshop is to describe a software release of our 'SciKnowMine' Document Triage System as a catalyst to finding an open-source infrastructure solution that powerfully frames this interaction between communities. The SciKnowMine architecture serves as an organizational framework for NLP specialists to frame experiments on standardized datasets as well as a delivery vehicle for their work so that it may be tested, scaled up and deployed within the curation workflows of biological databases. I will describe the software itself as well as a vision for how we hope to export this approach to deliver both text mining capabilites to biological scientists and important, impactful text-driven challenges to computer scientists.

Mouse Genome Informatics (MGI): the challenge to develop NLP strategies and tools to support publication-based curation

Janan Eppig, Mouse Genome Informatics, The Jackson Laboratory

The Mouse Genome Informatics (MGI, www.informatics.jax.org) Database supports biological knowledge building for the laboratory mouse, a species that is widely used as an animal model for studies of normal human biology and disease. MGI places particular emphasis on data integration to enable traversing and mining data from sequence-to-phenotype-to-human disease. Data include gene characterization and function, gene homologs, embryonic gene expression, nucleotide and protein sequences, SNPs, and mutant phenotypes and human disease associations. Data integration and standardization make it possible to use MGI as a tool for analyzing data relationships and for developing novel hypotheses. The majority of MGI data comes from large data providers, experimental consortia, and other electronically accessible sources. Application of standard nomenclatures and ontologies then enables data connections and integration to support discovery. A less automated, but very important data component of MGI includes the curation of the scientific literature. Peer reviewed publications continue (so far) to contain extensive and granular biological information and mechanistic studies generally not addressed in high- throughput projects. To bring this significant, heterogeneous body of information into semantic alignment and integrate it with MGI requires time-consuming work by highly trained biologists who are capable of reading, analyzing, and interpreting these data. The first steps: 1) locating MGI-relevant publications, 2) classifying publication data content (function, expression, disease model, etc.), and 3) indexing publications to the genomic features they describe (genes, mutations, variants, etc.) are the least complex, but time-intensive, and crucially necessary to the curation process. These steps appear most amenable to NLP applications and practically designed software tools. To this end, the SciKnowMine project has the potential to provide high cost-benefit to MGI and other publication-curated database resources.

The Colorado Richly Annotated Full Text corpus: A corpus linguist's perspective

Kevin Bretonnel Cohen, Computational Bioscience Program, University of Colorado.

Corpus linguistics is the science of building annotated text collections and using them for linguistic research. Building corpora always involves making mistakes, cutting corners, and other miscarriages of intent that affect the final product. This talk presents the Colorado Richly Annotated Full Text corpus from the point of view of a corpus linguist, with lessons learned from the process of building it.

Reactome – Linking pathways, networks and disease.

Robin Haw, Ontario Institute for Cancer Research, Informatics and Bio-Computing, Toronto, ON, Canada

Reactome is an open access, curated and peer-reviewed knowledgebase of human biological pathways and processes. The Reactome curation system draws upon the expertise of independent researchers who author precise machine-readable descriptions of human pathways under the guidance of a team of curators. Pathway modules are extensively checked to ensure factual accuracy and compliance with the data model, and a system of evidence tracking ensures that all assertions are backed by the primary literature. Recent extensions of our data model accommodate the annotation of disease processes, allowing us to represent the altered biological behavior of mutant variants frequently found in cancer, and to describe the mode of action and specificity of anti-cancer therapeutics. Reactome pathways currently cover a quarter of the translated portion of the genome, and are available on our web site for browsing, downloading, and manipulation by in-house and third party data analysis and visualization tools.

Database URL: http://www.reactome.org

OntoPub: An Ontology-Driven, Concept-Based Biomedical Literature Search Engine

Weisong Liu, Rat Genome Database

Biomedical researchers rely on NCBI's keyword-based search engine to locate articles of interest in the PubMed database. However, biomedical nomenclatures are not easily expressed in keywords. Synonyms and aliases often make it complicated to construct queries properly. At the Rat Genome Database, we built OntoPub by using ontologies, BioNLP tools and industrial level information processing tools. Free text in articles is tagged before being indexed. Query conditions in OntoPub consist of concepts defined in ontologies, biological entities such as genes and mutations, as well as keywords. Query expansion is fully automated which extremely simplifies the query process. OntoPub currently has been partially integrated into RGD's curation workflow. The query result in the curation tool lists abstracts with associated ontology and entity terms. Ontology terms and references can be selected for curation directly from the abstract list by the use of linked icons.

Wiki-based community annotation

Jim Hu, Biochemistry & Biophysics Department, Texas A&M University.

The success of Wikipedia has inspired a number of efforts to use collaborative editing tools to support community curation of biological databases. EcoliWiki and GONUTS (the Gene Ontology Normal Usage Tracking System) represent two examples of wiki-based biology resources that are maintained by my group. I will describe our approaches to address two challenges for community curation with wikis: 1) the wiki-specific problem of having structured data in a free text software platform and 2) the more general problem of promoting participation from the broader community. Structure in our wikis is currently provided by a combination of organizing pages using Mediawiki's category system and TableEdit, a system for working with tabular data. To address the problem of promoting community participation in curation, we are coupling curation with undergraduate education. Our pilot project for this is the Community Assessment of Community Annotation with Ontologies (CACAO), where GO annotation is done as an intercollegiate competition. NLP tools and integration with the semantic web will be important in promoting community curation by scientists and students who are domain experts but not professional biocurators.

An NLP Voyage: Explorations with Information Extraction for Biocuration

Ellen Riloff, School of Computing, University of Utah.

Many natural language processing (NLP) techniques have been developed to extract facts from text. These techniques can be used for application tasks that require fact extraction (e.g., identifying entities of a specific type) as well as classification tasks that can benefit from richer features. We will describe explorations with information extraction (IE) techniques for two biocuration tasks. First, we incorporated four types of contextual pattern features in a document triage classifier. We obtained performance gains by adding these features to an N-gram document classifier. Second, we employed a semantic dictionary learner to generate term lists for three semantic categories: cells, organs, and species. We applied this algorithm to journal articles of interest to biocurators at MGI. We will show how this IE technology can automatically produce phrase lists and contextual patterns to identify semantic entities that are relevant for biocuration.

Mining Fulltext

Maximilian Haeussler, UCSC Genome Browser Group.

Fulltext of research documents allows more information to be mined. I present an overview of how the genome browser at UCSC processes millions of fulltext files obtained from publishers, the types of information we extract and some document classification ideas we have recently implemented.

Clarifying ClearTK: A Case Study in Entity Tagging

James Gung, CLEAR group, University of Colorado Boulder

ClearTK is a framework for developing NLP components in Java built on top of Apache UIMA. In addition to providing a common interface and wrappers for many popular machine learning libraries, ClearTK includes an extensive feature extraction library easily used with the ML classifiers. ClearTK provides the infrastructure for tackling a wide variety of NLP tasks such as sequence tagging, relation extraction, summarization, and document classification. In this talk, I will discuss ClearTK in the context of a common NLP task, named entity recognition, thus introducing many of ClearTK's features and capabilities.

New Directions for Industry / Academic Collaborations in Knowledge Mining of Research Literature

William Gunn, Mendeley

This presentation will discuss ongoing work at Mendeley in recommendations, annotation, and entity extraction from research literature. The challenges we have encountered in creating and indexing documents and making structured representations of them will be discussed and I will talk about the emerging opportunities we have identified. I'll demonstrate some of the work we have done as part of the CODE project in the EU and then I'd like to have a discussion of the challenges in industry / academic collaborations.

BioCreative: Text Mining for Biocuration

Cecilia Arighi, Center for Bioinformatics and Computational Biology, University of Delaware, DE, USA

BioCreative: Critical Assessment of Information Extraction in Biology is an international community-wide effort that evaluates text mining (TM) and information extraction systems applied to the biomedical domain (http://www.biocreative.org/). A unique characteristic of this effort is its collaborative and interdisciplinary nature, as it brings together experts from various fields, including TM, biocuration, publishing houses and bioinformatics. Therefore each competition is tailored towards specific needs of these communities. In particular, BioCreative has been working closely with biocurators to understand the various curation workflows, the text mining tools that are being used and their major needs. In BioCreative workshop 2012 we reviewed descriptions of curation workflows from expert curated databases to identify commonalities and differences among these. One common theme was the need of semi- or fully-automated Gene Ontology (GO) curation techniques to assist database curators to rapidly identify relevant articles for GO curation, and as a result, a track for this topic has been included in the BioCreative IV competition.

To address the current barriers in using text mining in biology, BioCreative has further been conducting user requirements analysis, user-based evaluations and fostering standards development for text mining tool re-use and integration. In this respect, the BioCreative Interactive Text Mining (IAT) task has served as a great means to observe the approaches, standards and functionalities used by state-of-the-art text mining systems with potential applications in the biocuration domain. The IAT task also provides a means for biocurators to be directly involved in the testing of text mining systems. The benefits to biocurators participating in this activity are multifold, including: direct communication and interaction with developers; exposure to new text mining tools that can be potentially adapted and integrated into the biocuration workflow, contribution to the development of text mining systems that meet the needs of the biocuration community, and dissemination of findings in peer-reviewed journal articles. A User Advisory Group (UAG) representing a diverse group of users with literature-based curation needs has been assisting in the design and assessment of the IAT and other tracks. This talk will present an overview of the BioCreative workshops with emphasis on the efforts that involve direct interaction with the biocuration community.

Speaker Contact Information

Cecilia Algerheri	arighi@dbi.udel.edu
Judy Blake	judith.blake@jax.org
Kevin Cohen	kevin.cohen@gmail.com
Harold Drabkin	harold.drabkin@jax.org
Janan Eppig	Janan.Eppig@jax.org
Jack Gardiner	jack.m.gardiner@gmail.com
James Gung	gungjm@gmail.com
William Gunn	william.gunn@mendeley.com
Jim Hu	jimhu@tamu.edu
Maximilian Haeussler	max@soe.ucsc.edu
Robin Haw	robin.haw@oicr.on.ca
Jim Kadin	James.Kadin@jax.org
Donghui Li	donghui@stanford.edu
Weisong Liu	wliu@mcw.edu
Paul Lloyd	plloyd@stanford.edu
Ellen Riloff	riloff@cs.utah.edu