Pacific Symposium on Biocomputing 2006: World Scientific; 2006. p. 1-3.

LINKING BIOMEDICAL INFORMATION THROUGH TEXT MINING: SESSION INTRODUCTION

KEVIN BRETONNEL COHEN
Center for Computational Pharmacology
University of Colorado
Denver, CO 80045, USA

OLIVIER BODENREIDER
National Library of Medicine
Bethesda, MD 20894, USA

LYNETTE HIRSCHMAN
The MITRE Corporation
Bedford, MA 01730, USA

This session is focused on text mining applications that link information from the biomedical literature to the growing array of structured resources available to researchers, such as protein databases (e.g., UniProt, PDB, PIR), model organism databases (e.g., FlyBase, MGI, SGD), ontologies (the Gene Ontology, as well as the growing number of ontologies in OBO – Open Biological Ontologies), and nomenclatures (HUGO, HUPO). To achieve this focus, there was an explicit requirement that submissions include both a text mining component and a mapping between at least two publicly available data sources. There were twenty papers submitted to this session, with nine papers accepted (7 for oral presentation).

This session builds on two threads of work that have been well represented at past PSB meetings, namely text mining and ontologies. There have been PSB sessions on text mining in 2000, 2001, 2002, and 2003. Many of the systems discussed in these earlier sessions focused on recognition of biomedical entities and relations in order to provide effective indexing into the literature. Other papers focused on topic-based document clustering, to provide tools to manage the vast biomedical literature at the document level. However, these systems were limited in that they did not link to resources outside the text collections (generally PubMed). The entity recognition systems identified entities or relations by simply pointing to substrings in the input

text. Such outputs are of intrinsically limited value. For example, a system that produces a table of protein-protein interactions is potentially highly valuable if it refers to specific entities in PDB, but of much more limited utility if it outputs only a list of potentially ambiguous symbols and names.

The second relevant thread at PSB investigated the linguistic and semantic characteristics of a variety of publicly available biomedical data sources, including gene names and Gene Ontology terms. Much of this work was presented at PSB sessions on ontologies in 2003, 2004, and 2005 or the PSB sessions on biomedical language processing listed above. Of particular interest is the identification of various kinds of relations among biomedical entities, which can enrich existing ontologies and subsequently benefit text mining.

This 2006 session on linking biomedical information represents the logical next step. Our goal has been to solicit papers that follow through on the insights gained into the structure of available data sources and advances in text mining, to create language processing systems that not only locate information in texts, but also map it to these explicit knowledge models. Two recent competitive evaluation tasks from BioCreAtIvE (Critical Assessment of Information Extraction in Biology) showed that it is possible to create systems that produce grounded outputs and to perform principled evaluations of them. BioCreAtIvE Task 1b [1] involved mapping references to genes in abstracts to specific gene identifiers from the appropriate model organism database. BioCreAtIvE Task 2 [2] involved assigning Gene Ontology terms to proteins mentioned in journal articles. Taken together, these two tasks demonstrate that it is possible to link the literature to specific entities and to specific concepts. At the same time, they make it clear that there is considerable room for improvement in performance of these tasks.

The papers for this session demonstrate the progress that has been made in using text mining to link across resources and to anchor mentions of biological entities to accepted biological nomenclatures and ontologies. These papers tackle a number of biological problems using a variety of technologies:

- Four papers emphasize the linkage to ontologies. One paper (Johnson et al.) discusses lexically-based techniques for ontology alignment between GO and several other ontologies. The other three papers focus on annotation into an ontology: Höglund et al. produce improved results for subcellular localization by combining both sequence data and text mining; Lussier et al. describes

PhenoGO, a system that maps from text into one of several anatomical ontologies; and Stoica and Hearst describe improved results for BioCreAtIvE Task 2 (functional annotation of papers on human proteins) by using orthologous genes in Mouse.

- Two papers describe summarization applications: Lu et al. focus on generation of GeneRIFs based on overlap of GO annotations with PubMed abstracts; Ling et al. describe an algorithm for generation of summaries by identifying documents about a particular gene and then extracting the most relevant sentence(s) for six aspects of gene function to create the summary.
- The paper by Vlachos et al. uses relations from the Sequence Ontology to improve on named entity results for FlyBase genes and to support an ontology-based coreference resolution strategy for these genes and gene products.

Overall, the papers in this session reflect the growing maturity of text mining as a bioinformatics tool that can be used, often in conjuction with other bioinformatics tools, to extract knowledge from the biomedical literature and to integrate it effectively with other knowledge sources.

References

1. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics. 2005;6 Suppl 1:S11.
2. Blaschke C, Leon EA, Krallinger M, Valencia A. Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics. 2005;6 Suppl 1:S16.