Pacific Symposium on Biocomputing 13:551-555(2008)

TRANSLATING BIOLOGY: TEXT MINING TOOLS THAT WORK

K. BRETONNEL COHEN, HONG YU, PHILIP E. BOURNE, AND LYNETTE HIRSCHMAN*

1. Introduction

This year is the culmination of two series of sessions on natural language processing and text mining at the Pacific Symposium on Biocomputing. The first series of sessions, held in 2001, 2002, and 2003, explored information extraction and retrieval applications for a range of possible biomedical applications. The second series of sessions began in 2006. In the first two years of this series, the sessions focused on tasks that required mapping to or between grounded entities in databases (2006) and on cutting-edge problems in the field (2007). The goal of this final session of the second series has been to assess where the past several years' worth of work have gotten us, what sorts of deployed systems have resulted, how well the systems have integrated genomic databases and the biomedical literature, and how usable these systems are. To this end, we solicited papers that addressed the following questions:

- What is the actual utility of text mining in the workflows of the various communities of potential users—model organism database curators, bedside clinicians, biologists utilizing high-throughput experimental assays, hospital billing departments, etc.?
- How usable are biomedical text mining applications? How does the application fit into the workflow of a complex bioinformatics

*KBC: Center for Computational Pharmacology; work supported by NIH grants R01-LM008111 and R01-LM009254 to Lawrence Hunter. HY: U. of Wisconsin-Madison; work supported by U. of Wisconsin Research Committee Awards, Research Growth Initiative grants, and an MiTAG award, and NIH grant R01-LM009836-01A1. PEB: UCSD, PDB. LH: The MITRE Corporation. Work supported by grant II-0640153 from the US National Science Foundation.

pipeline? What kind of training does a bioscientist require to be able to use an application?

- Is it possible to build portable text mining systems? Can systems be adapted to specific domains and specific tasks without the assistance of an experienced language processing specialist?
- How robust and reliable are biomedical text mining applications? What are the best ways to assess robustness and reliability? Are the standard evaluation paradigms of the natural language processing world—intrinsic evaluation against a gold standard, post-hoc judging of outputs by trained judges, extrinsic evaluation in the context of some other task—the best evaluation paradigms for biomedical text mining, or even sufficient evaluation paradigms?

2. The session

We received 29 submissions and accepted nine papers. Each paper received at least three reviews by members of a program committee composed of biomedical language processing specialists and computational biologists from North America, Europe, and Asia. All four of the broad questions were addressed by at least one paper. We review all nine papers briefly here.

Utility: A number of papers addressed the issue of utility. Alex et al.[1] experimented with a variety of forms of automated curator assistance, measuring curation time and assessing curator attitudes by questionnaire, and found that text mining techniques can reduce curation times by as much as one third. Caporaso et al.[3] examined potential roles for text-based and alignment-based methods of annotating mutations in a database curation workflow. They found that text mining techniques can provide a quality assurance mechanism for genomic databases. Roberts and Hayes[9] analyzed a large collection of information requests from an understudied population—commercial drug developers—and found that various families of text mining solutions can play a role in meeting the information needs of this group. Wang et al. [11] evaluated a variety of algorithms for gene normalization, and found that there are complex interactions between performance on a gold standard, improvement in curator efficiency, portability, and the demands of different kinds of curation tasks.

Usability: Divoli et al.[4] applied a user-centered design methodology to investigate the kinds of information that users want to see displayed in interfaces for performing biomedical literature searches. Among other

findings, they report that users showed interest in having gene synonyms displayed as part of the search interface, and that they would like to see extracted information about genes, such as chemicals and drugs with which they are associated, displayed as part of the results.

Portability: Leaman and Gonzalez[8] focused on portability of gene mention detection techniques across semantic classes of named entities and across corpora. Wang et al.[11] examined portability issues in their study of the effects of various gene normalization algorithms on curator efficiency. However, the challenge of building systems that can be ported to new domains without the assistance of a text mining specialist remains unaddressed.

Robustness and reliability: Several papers looked at the adequacy of traditional text mining evaluation paradigms, either directly or indirectly. Caporaso et al.[3] examined the correspondence between system performance on intrinsic and extrinsic evaluations, and found that high performance on a corpus does not necessarily predict high performance on an actual annotation task, due in part to the necessity of access to full-text journal articles for database curation. Kano et al.[7] explored the role of well-engineered integration platforms in building complex language processing systems from independent components, and showed that a well-designed platform can be used to determine the optimum set of components to combine for a specific relation extraction task. Wang et al.[11] found that the best-performing algorithms for gene normalization as determined by intrinsic evaluation against a gold-standard data set is not necessarily the most effective algorithm for accelerating curation time.

Other topics: Dudley and Butte[5] explored the use of simple pattern-matching techniques to solve a fundamental problem in translational medicine: finding expression array data sets that pair disease-related experimental conditions with those from normal controls. This paper illustrates the power of mining large data collections with simple tools to extract high-value data sets. Finally, Brady and Shatkay[2] demonstrated that text mining can be used to apply subcellular localization prediction to almost any protein, even in the absence of published data about it.

3. Conclusions

Some of the most influential and frequently-cited papers in what might be called the "genomic era" of biomedical language processing were presented at PSB. Fukuda et al.'s early and oft-cited paper on named entity recogni-

tion for the gene mention problem[6] appeared at PSB in 1998; more recently, Schwartz and Hearst's algorithm for identifying abbreviation definitions in biomedical text[10] rapidly became one of the most frequently used components of biomedical text mining systems after being presented at PSB in 2003. The years since the first PSB text mining sessions have seen phenomenal growth in the work on biomedical text mining, including several deployed systems, commercial tools, systematic challenge evaluations, and an expansion of text mining into the computational biology workflow. The work presented in this year's session suggests that we are now poised to tap the potential of text mining to contribute to mainstream computational bioscience.

References

1. B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, and X. Wang. Assisted curation: Does text mining really help? In *Pac Symp Biocomput*, 2008.
2. S. Brady and H. Shatkay. EpiLoc: A (working) text-based system for predicting protein subcellular location. In *Pac Symp Biocomput*, 2008.
3. J. G. Caporaso, N. Deshpande, J. L. Fink, P. E. Bourne, K. B. Cohen, and L. Hunter. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. In *Pac Symp Biocomput*, 2008.
4. A. Divoli, M. A. Hearst, and M. A. Wooldridge. Evidence for showing gene/protein name suggestions in bioscience literature search interfaces. In *Pac Symp Biocomput*, 2008.
5. J. Dudley and A. J. Butte. Enabling integrative genomic analysis of high-impact human diseases through text mining. In *Pac Symp Biocomput*, 2008.
6. K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi. Toward information extraction: identifying protein names from biological papers. In *Pac Symp Biocomput*, pages 707–718, 1998.
7. Y. Kano, N. Nguyen, R. Sætre, K. Yoshida, Y. Miyao, Y. Tsuruoka, Y. Matsubayashi, S. Ananiadou, and J. Tsujii. Filling the gaps between tools and users: A tool comparator, using protein-protein interaction as an example. In *Pac Symp Biocomput*, 2008.
8. R. Leaman and G. Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition. In *Pac Symp Biocomput*, 2008.
9. P. M. Roberts and W. S. Hayes. Information needs and the role of text mining in drug development. In *Pac Symp Biocomput*, 2008.
10. A. Schwartz and M. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pac Symp Biocomput*, volume 8, pages 451–462, 2003.
11. X. Wang and M. Matthews. Comparing usability of matching techniques for normalising biomedical named entities. In *Pac Symp Biocomput*, 2008.