Pacific Symposium on Biocomputing 12:205-208(2007)

NEW FRONTIERS IN BIOMEDICAL TEXT MINING

PIERRE ZWEIGENBAUM, DINA DEMNER-FUSHMAN, HONG YU, AND K. BRETONNEL COHEN

1. Introduction

To paraphrase Gildea and Jurafsky [7], the past few years have been exhilarating ones for biomedical language processing. In less than a decade, we have seen an amazing increase in activity in text mining in the genomic domain [20]. The first textbook on biomedical text mining with a strong genomics focus appeared in 2005 [3]. The following year saw the establishment of a national center for text mining under the leadership of committed members of the BioNLP world [2], and two shared tasks [10,9] have led to the creation of new datasets and a very large community.

These years have included considerable progress in some areas. The TREC Genomics track has brought an unprecedented amount of attention to the domain of biomedical information retrieval [8] and related tasks such as document classification [5] and question-answering, and the BioCreative shared task did the same for genomic named entity recognition, entity normalization, and information extraction [10].

Recent meetings have pushed the focus of biomedical NLP into new areas. A session at the Pacific Symposium on Biocomputing (PSB) 2006 [6] focussed on systems that linked multiple biological data sources, and the BioNLP'06 meeting [20] focussed on deeper semantic relations. However, there remain many application areas and approaches in which there is still an enormous amount of work to be done.

In an attempt to facilitate movement of the field in those directions, the Call for Papers for this year's PSB natural language processing session was written to address some of the potential "New Frontiers" in biomedical text mining. We solicited work in these specific areas:

- Question-answering
- Summarization
- Mining data from full text, including figures and tables
- Coreference resolution

- User-driven systems
- Evaluation

31 submissions were received. Each paper received four reviews by a program committee composed of biomedical language processing specialists from North America, Europe, and Asia. Eleven papers were selected for publication. The papers published here present an interesting window on the nature of the frontier, both in terms of how far it has advanced, and in terms of which of its borders it will be difficult to cross.

One paper addresses the topic of summarization. Lu et al. [14] use summary revision techniques to address quality assurance issues in GeneRIFs.

Two papers extend the reach of biomedical text mining from the abstracts that have been the input to most BioNLP systems to date, towards mining the information present in full-text journal articles. Kou et al. [13] introduce a method for matching the labels of sub-figures with sentences in the paper. Seki and Mostafa [19] explore the use of full text in discovering information not explicitly stated in the text.

Two papers address the all-too-often-neglected issue of the usability and utility of text mining systems. Karamanis et al. [12] present an unusual attempt to evaluate the usability of a system built for model organism database curators. Much of the work in biomedical language processing in recent years has assumed the model organism database curator as its user, so usability studies are well-motivated. Yu and Kaufman [22] examine the usability of four different biomedical question-answering systems.

Two papers fit clearly into the domain of evaluation. Morgan et al. [15] describe the design of a shared evaluation, and also gives valuable baseline data for the entity normalization task. Johnson et al. [11] describe a fault model for evaluating ontology matching, alignment, and linking systems.

Four papers addressed more traditional application types, but at a deeper level of semantic sophistication than most past work in their areas. Two papers dealt with the topic of relation extraction. Ahlers et al. [1] tackle an application area—information extraction—that has been a common topic of previous work in this domain, but does so at an unusual level of semantic sophistication. Cakmak and Özsoyoglu [4] deal with the difficult problem of Gene Ontology concept assignment to genes. Finally, two papers focus on the well-known task of document indexing, but at unusual levels of refinement. Névéol et al. [16] extract MeSH *sub*headings and pairs them with the appropriate primary heading, introducing an element of context that is lacking in most other work in BioNLP. Rhodes et al. [18]

describe a methodology for indexing documents based on the structure of chemicals that are mentioned within them.

So, we see papers in some of the traditional aplication areas, but at increased levels of sophistication; we see papers in the areas of summarization, full text, user-driven work, and evaluation; but no papers in the areas of coreference resolution or question-answering. What might explain these gaps? One possibility is the shortage of publicly available datasets for system building and evaluation. Although there has been substantial annotation work done in the area of coreference in the molecular biology domain [21,17], only a single biomedical corpus with coreference annotation is currently freely available [17]. Similarly, although the situation will be different a year from now due to the efforts of the TREC Genomics track, there are currently no datasets freely available for the biomedical question-answering task.

2. Acknowledgments

K. Bretonnel Cohen's participation in this work was supported by NIH grant R01-LM008111 to Lawrence Hunter.

References

1. Caroline B. Ahlers, Marcelo Fiszman, Dina Demner-Fushman, François Michel Lang, and Thomas C. Rindflesch. Extracting semantic predications from MEDLINE citations for pharmacogenomics. In *Pacific Symposium on Biocomputing*, 2007.
2. Sophia Ananiadou, Julia Chruszcz, John Keane, John McNaught, and Paul Watry. The National Centre for Text Mining: aims and objectives. *Ariadne*, 42, 2005.
3. Sophia Ananiadou and John McNaught. *Text mining for biology and biomedicine*. Artech House Publishers, 2005.
4. Ali Cakmak and Gultekin Özsoyoğlu. Annotating genes by mining PubMed. In *Pacific Symposium on Biocomputing*, 2007.
5. Aaron M. Cohen and William R. Hersh. The TREC 2004 genomics track categorization task: classifying full text biomedical documents. *Journal of Biomedical Discovery and Collaboration*, 1(4), 2006.
6. K. Bretonnel Cohen, Olivier Bodenreider, and Lynette Hirschman. Linking biomedical information through text mining: session introduction. In *Pacific Symposium on Biocomputing*, pages 1–3, 2006.
7. Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.
8. William R. Hersh, Ravi Teja Bhupatiraju, Laura Ross, Phoebe Roberts, Aaron M. Cohen, and Dale F. Kraemer. Enhancing access to the Bibliome:

the TREC 2004 Genomics track. *Journal of Biomedical Discovery and Collaboration*, 2006.

9. William R. Hersh, Aaron M. Cohen, Jianji Yang, Ravi Teja Bhupatiraju, Phoebe Roberts, and Marti Hearst. TREC 2005 Genomics track overview. In *Proceedings of the 14th Text Retrieval Conference*. National Institute of Standards and Technology, 2005.
10. Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6, 2005.
11. Helen L. Johnson, K. Bretonnel Cohen, and Lawrence Hunter. A fault model for ontology mapping, alignment, and linking systems. In *Pacific Symposium on Biocomputing*, 2007.
12. Nikiforos Karamanis, Ian Lewin, Ruth Seal, Rachel Drysdale, and Edward J. Briscoe. Integrating natural language processing with FlyBase curation. In *Pacific Symposium on Biocomputing*, 2007.
13. Zhenzhen Kou, William W. Cohen, and Robert F. Murphy. A stacked graphical model for associating information from text and images in figures. In *Pacific Symposium on Biocomputing*, 2007.
14. Zhiyong Lu, K. Bretonnel Cohen, and Lawrence Hunter. GeneRIF quality assurance as summary revision. In *Pacific Symposium on Biocomputing*, 2007.
15. Alexander A. Morgan, Benjamin Wellner, Jeffrey B. Colombe, Robert Arens, Marc E. Colosimo, and Lynette Hirschman. Evaluating human gene and protein mention normalization to unique identifiers. In *Pacific Symposium on Biocomputing*, 2007.
16. Aurélie Névéol, Sonya E. Shooshan, Susanne M. Humphrey, Thomas C. Rindflesch, and Alan R. Aronson. Multiple approaches to fine indexing of the biomedical literature. In *Pacific Symposium on Biocomputing*, 2007.
17. J. Pustejovsky, J. Castaño, R. Saurí, J. Zhang, and W. Luo. Medstract: creating large-scale information servers for biomedical libraries. In *Natural language processing in the biomedical domain*, pages 85–92. Association for Computational Linguistics, 2002.
18. James Rhodes, Stephen Boyer, Jeffrey Kreulen, Ying Chen, and Patricia Ordonez. Mining patents using molecular similarity search. In *Pacific Symposium on Biocomputing*, 2007.
19. Kazuhiro Seki and Javed Mostafa. Discovering implicit associations beween genes and hereditary diseases. In *Pacific Symposium on Biocomputing*, 2007.
20. Karin Verspoor, K. Bretonnel Cohen, Inderjeet Mani, and Benjamin Goertzel. Introduction to BioNLP'06. In *Linking natural language processing and biology: towards deeper biological literature analysis*, pages iii–iv. Association for Computational Linguistics, 2006.
21. Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. Improving noun phrase coreference resolution by matching strings. In *IJCNLP04*, pages 326–333, 2004.
22. Hong Yu and David Kaufman. A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. In *Pacific Symposium on Biocomputing*, 2007.