Journal of Biomedical Informatics 42 (2009) 757–759

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Guest Editorial

Current issues in biomedical text mining and natural language processing

The years since 1998 have seen an explosion in work in biomedical text mining (BioNLP) of both clinical text and the biomedical literature [1]. The work focusing on the literature has been particularly stimulated by three factors. One is simply the rapid increase in the rate of publication in general, as reflected in the growth in the contents of PubMed/MEDLINE, which has been exponential. Another is the growth in the use of high-throughput assays, which commonly produce lists of genes much larger than were seen in previous experimental methods. Interpreting these gene lists typically requires the digestion of large amounts of published literature. Finally, the construction of model organism and other databases has been unable to keep up with the rate of discovery of the entities that they describe [2]. Some have observed that BioNLP, serving as a curator aid, is a potential solution to this problem.

In the clinical domain, there has been a surge of interest in using electronic medical record (EMR) systems to improve the quality of care through decision support, evidence-based medicine, and outbreak and disease surveillance. To make full use of the information contained in EMRs, we have had to tackle the mass of textual data that make up a large portion of the patient record. Thus, there seems to be an increased demand for information retrieval and extraction tools for clinical narratives. In this special issue we begin with a methodological review titled "What Can Natural Language Processing do for Clinical Decision Support?" in which Demner-Fushman and colleagues examine the role NLP in point-of-care decision support [3]. They discuss the evolution of clinical NLP from early innovation to stable research at major clinical centers to a shift toward mainstream interest in biomedical NLP. Motivated by the demand for clinical language processing, the BioNLP community has shown renewed interest in the development of fundamental NLP methods and advances in NLP systems for clinical decision support.

Another contributor to the growth of BioNLP has been the availability of a wide range of resources suitable for analysis or tools for assisting in such analyses. These include no-cost textual sources such as PubMed/MEDLINE and PubMedCentral; a large variety of corpora [4]; ontologies and other lexical semantic resources, such as the Gene Ontology, UMLS, and the Semantic Network; databases such as Entrez Gene and DIP; and NLP pipelines for clinical information extraction [5,6]. These resources have provided grist for the processing mill, reference resources, shareable tools, and—in the case of the databases—both lexical resources and sources of gold-standard data for tasks like protein–protein interaction. Resources have tended to be more plentiful in the areas of biology and biomedical literature, but a few shared tasks [7,8] have made available de-identified clinical data sets. A new resource for clinical NLP is the University of Pittsburgh's NLP Repository of one year's worth of de-identified clinical reports for NLP research [9]. We suspect that as resources continue to grow, BioNLP research will continue to expand, and performance will increase.

Finally, government funding for work in this area has been an important stimulus to BioNLP research.

One sign of the growth of work in BioNLP is this special issue, which garnered the largest number of submissions (33) that JBI has ever received in response to a single call for papers. The submissions broadly covered the field of BioNLP—from clinical to biological domains, from literature to clinical text genres, from knowledge-based to purely statistical methods, and from basic named-entity recognition to sophisticated analyses. Broad categories of work include document classification, named-entity recognition, ontology development, information extraction and coding, segmentation, genre analysis indexing, and summarization. We give here a brief description of the 19 accepted research papers (we also feature the one methodological review that was mentioned above); our summary itself gives a snapshot of the state of BioNLP today.

Biomedical NLP is expanding its reach to a variety of biomedical texts. Consistent with recent trends in biomedical NLP, over half of the articles (11 of 19) process the biomedical literature, and a quarter process physician-dictated clinical reports (5). NLP systems are reaching beyond these traditional data sources to an assortment of textual data, including disease outbreak reports [10], medical student clinical notes [11], and reports of randomized controlled trials [12].

One reason NLP systems are being developed for data sources beyond the literature and clinical reports is that many of the systems described in this special issue are application-driven—specific biomedical applications require information from text, and those requirements drive the NLP development. Applications driving development of methodologies described in this issue tend to be tailored toward semi-automated tools for aiding human beings in particular tasks.

For instance, in the first research paper in this issue, Fiszman and colleagues describe an automatic graphical summarization system (Semantic MEDLINE) to help physicians and scientists navigate the vast amount of information described in MEDLINE citations [13]. They have evaluated the summarization system's ability to identify useful drug interventions for 53 diseases, and they show that the system increased both mean average precision and clinical usefulness over a baseline system. Similarly, in the next paper, Neveol and colleagues identify entities in MEDLINE citations, but with the goal of assisting human indexers in selecting the best MeSH heading/subheading pairs, using the National Library of Medicine's Medical Text Indexer [14]. They experiment with combinations of NLP, statistical, and machine learning

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.09.001

techniques and measure which subheading attachments were selected by human indexers.

In the next paper, Denny and colleagues apply the KnowledgeMap system to the task of tracking medical students' clinical experiences. They showed that it was possible to identify the extent to which students were exposed to core clinical problems by identifying UMLS concepts in the students' dictated notes, providing a recall of 0.91 and precision of 0.92 without requiring additional work from medical trainees [11]. Doan and colleagues then describe their experiment with the text mining component of the BioCaster public health protection system, which aids humans in tracking and understanding disease outbreaks across the globe [10]. They have evaluated a text classifier for distinguishing between general disease-oriented news and infectious disease outbreak reports on the Internet and show improved performance over raw text when including features that take into account roles in combination with both named entities and semantic categories of disease-related nouns and verbs.

Aiding humans in developing ontologies by extracting potential candidate terms from published literature is the motivation for the next article, by Zheng and colleagues [15]. The authors seek to prioritize candidate terms objectively by quantifying each term's relevance to the domain within the biomedical literature. Using two biomedical domains—the Gene Ontology and a Clinical Trial Ontology—they apply a computational method that utilizes a text mining approach based on the hyper-geometric enrichment test and demonstrate that a term's over-representation in domain-related PubMed abstracts is an indication of that term's relevance to the domain.

Like presentations in the AMIA tracks in foundational informatics and the Association for Computational Linguistics' SIGBIOMED-associated meetings, the articles in this issue not only demonstrate methods for successfully addressing the goal of a specified application but also explore the foundational underpinnings of data sources and techniques, with hopes that understanding the underlying structure and characteristics of the task will ultimately lead to better modeling and improved performance of automated techniques. Studies address questions such as "How often do users address more than one subject in a PubMed query," "How are historical findings exhibited in different types of clinical reports," and "How are the intervention arms of a trial expressed syntactically."

For instance, we turn to a paper by Lu and Wilbur in which they investigate the problem of grouping user queries in PubMed with the goal of facilitating studies of searching behavior [16]. They apply lexical and contextual analyses, achieving an accuracy of 90.7%, and find that a significant proportion of PubMed queries involve multiple, lexically-related queries. The subsequent paper, by Harkema and colleagues, investigates the portability of the ConText algorithm, an extension of NegEx, across various genres of clinical reports [17]. For six report genres the authors measured the prevalence of clinical conditions expressed as absent, hypothetical, historical, and experienced by someone other than the patient. They found that distributions of these contextual features were extremely varied and that the algorithm did not perform as well on some report genres. In the next paper, Chung investigates the role of coordinating constructions in describing the intervention arms of clinical trials [12], finding that interventions are most often described in coordinating constructions and developing a method for identifying coordinating constructions that performs with an *F*-measure of 0.78 using full syntactic parsing, predicate-argument structure, and other linguistic features. Then Rimell and Clark investigate parser retraining on biomedical text [18], showing that lexicalized grammar formalisms such as Combinatorial Categorial Grammar may allow for more successful domain-specific retraining than has been reported in previous work.

Foundational research is critical to boosting performance of NLP techniques beyond that obtained through generic approaches—if we want to advance the performance of our systems beyond the "80/20" often obtained through initial machine learning experiments, we need a deep understanding of the syntactic, semantic, and discourse characteristics of the data we are processing.

Development of NLP applications in the diverse areas represented in this special issue necessitates an extensive array of natural language processing methodologies, and authors of these articles illustrate successful examples of syntactic parsing, textual summarization, single and multi-class classification, segmentation, named-entity recognition, relation extraction, indexing to a standardized vocabulary, and encoding concepts and their characteristics. To accomplish these tasks, authors used wide-ranging techniques that take advantage of domain, statistical, and linguistic information. Although some studies focus on only one technique, the majority integrate multiple methods to accomplish their aims. Moving beyond the simple but successful n-gram representation of text for statistical classification, the objective of several studies was to determine which combination of linguistic and/or domain-specific features improved classification performance. These experiments leverage the power of statistical pattern matching while also integrating linguistic characteristics and expert knowledge, the combination of which we believe is essential to the holy grail of biomedical natural language understanding.

A number of papers deal with document-level classification. In the next paper, for example, Lan and colleagues investigate the problem of finding documents relevant to protein–protein interactions, using the BioCreative II dataset [19]. They investigated a number of novel and domain-specific features and improved performance over previous methods. Similarly, Goldstein and Uzuner's paper investigates the classification of discharge summaries with respect to whether they mention obesity or one of its comorbidities [20]. They evaluate an approach based on specialized per-disease classifiers, comparing it to voting and stacking approaches, and find that it outperforms both of them.

The next set of papers deals with the named-entity recognition task. Nenadic and colleagues attempt the novel task of identifying only proteins that are transcription factors [21]. They achieve an *F*-measure of 0.52 on fivefold cross-validation, demonstrating the difficulty of the task of assigning specific semantic classes to the already constrained class of proteins. In the next study, Smith and Wilbur evaluate the contribution of seven different parsers to the task of determining whether a base noun phrase contains a gene mention, using the GENETAG corpus [22]. They show that information from the parsers improves performance on the task but that differences among parsers are negligible. Then Saha and colleagues investigate named-entity recognition for several biological semantic classes, using the JNLPBA data [23]. They explore the use of word clustering and selection techniques for this task and show that this approach could outperform other knowledge-free approaches. In the next article, Hsiao and colleagues use the UMLS to construct a hierarchical vocabulary of concepts related to brain function and experimental methods, developing a sophisticated named-entity recognition system for those concepts [24]. Their approach not only recognizes mentions of terms related to brain function and experimental methods, but normalizes them to concepts in the hierarchical vocabulary.

In the next set of papers, authors deal with information extraction and encoding. In the first study, Mykowiecka and colleagues discuss the development of systems for extraction of medical data about mammography reports and diabetes from Polish-language clinical documents [25]. The work is unusual both for its input language and for its completely rule-based effort to structure the task using an ontology. Then Coden and colleagues report on the use of an ontology, specialized for the cancer domain [26]. They

developed a system for semantic class recognition and relation extraction to the knowledge model, achieving a variety of performance levels that vary with the prevalence of the target in the input texts.

Finally, to close the issue, a small set of papers deals with corpora and annotation. First, Roberts and colleagues report on the construction of a corpus built from sampled text from 20,000 clinical records [27]. Annotation guidelines are given, along with lessons learned from the process. Then Peshkin and colleagues describe an annotation schema for binary relations, a publicly available distributed annotation tool, and a pilot annotation effort on autism [28]. Inter-annotator agreement of 75% is reported, suggesting that the annotation schema and approach are viable.

For all of the diversity of domains, genres, methods, tasks, and applications that we see represented in this special issue, there are still some areas that are missing. Cohen et al. [29] identified four application areas that have been underrepresented or absent in BioNLP research: usability, utility, portability, and robustness and reliability. We note that these are similarly mostly absent from the papers in this issue, although one article [13] presents a novel clinical usefulness metric. The topic of portability appears, but only in the sense of portability across genres [17]—there is no discussion of systems that can be ported by non-text-mining-expert users. As BioNLP systems evolve from the lab to use in real settings, we expect to see more investigations addressing these critical measures of success.

References

[1] Verspoor K, Bretonnel Cohen K, Goertzel K, Mani I. Linking natural language processing and biology: towards deeper biological literature analysis. In: BioNLP '06: Proceedings of the workshop on linking natural language processing and biology 2006. Morristown, New Hersey, USA: Association for Computational Linguistics; 2006. p. iii-iv.

[2] Baumgartner Jr WA, Bretonnel Cohen K, Fox L, Acquaah-Mensah G, Hunter L. Manual annotation is not sufficient for curating genomic databases. Bioinformatics 2007;23:i41–8.

[3] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42(5):760–72.

[4] http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml [accessed 19.08.09].

[5] Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. AMIA Annu Symp Proc 2006:931.

[6] https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP_Documentation_and_Downloads [accessed 19.08.09].

[7] Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. AMIA Annu Symp Proc 2008:1252–3.

[8] Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, et al. A shared task involving multi-label classification of clinical free text. In: BioNLP '07: Proceedings of the workshop on BioNLP 2007. Morristown, NJ, USA: Association for Computational Linguistics; 2007. p. 97–104.

[9] http://www.dbmi.pitt.edu/blulab/nlprepository.html [accessed 19.08.09].

[10] Doan S, Kawazoe A, Conway M, Collier N. Towards role-based filtering of disease outbreak reports. J Biomed Inform 2009;42(5):773–80.

[11] Denny JC, Bastarache L, Sastre EA, Spickard III A. Tracking medical students' clinical experiences using natural language processing. J Biomed Inform 2009;42(5):781–9.

[12] Yuet-Chee Chung G. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. J Biomed Inform 2009;42(5):790–800.

[13] Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform 2009;42(5):801–13.

[14] Neveol A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR. A recent advance in the automatic indexing of the biomedical literature. J Biomed Inform 2009;42(5):814–23.

[15] Zheng Z, Tsoi LC, Patel R, Zhao W. Text mining approach to evaluate terms for ontology development. J Biomed Inform 2009;42(5):824–30.

[16] Lu Z, Wilbur WJ. Improving accuracy for identifying related PubMed queries by an integrated approach. J Biomed Inform 2009;42(5):831–8.

[17] Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009;42(5):839–51.

[18] Rimell L, Clark S. Porting a lexicalized grammar parser to the biomedical domain. J Biomed Inform 2009;42(5):852–65.

[19] Lan M, Tan CL, Su J. Feature generation and representations for protein–protein interaction classification. J Biomed Inform 2009;42(5):866–72.

[20] Goldstein I, Uzuner O. Specializing for predicting obesity and its co-morbidities. J Biomed Inform 2009;42(5):873–86.

[21] Nenadic G, Yang H, Keane J, Bergman CM. Assigning roles to protein mentions: the case of transcription factors. J Biomed Inform 2009;42(5):887–94.

[22] Smith L, Wilbur J. The value of parsing as feature generation for gene mention recognition. J Biomed Inform 2009;42(5):895–904.

[23] Saha SK, Sarkar S, Mitra P. Feature selection for maximum entropy based biomedical named entity recognition. J Biomed Inform 2009;42(5):905–11.

[24] Hsiao MY, Chen CC, Chen JH. Using UMLS to construct a generalized hierarchical concept-based dictionary of brain functions for information extraction from fMRI literature. J Biomed Inform 2009;42(5):912–22.

[25] Mykowiecka A, Marciniak M, Kupsc A. Rule-based information extraction from patients' clinical data. J Biomed Inform 2009;42(5):923–36.

[26] Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, deGroen PC. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge model. J Biomed Inform 2009;42(5):937–49.

[27] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform 2009;42(5):950–66.

[28] Peshkin L, Monaghan T, Blanco A, Wall DP, Cano C. Collaborative annotation resource for disease-centered relation extraction from biomedical text. J Biomed Inform 2009;42(5):967–77.

[29] Bretonnel Cohen K, Yu H, Bourne PE, Hirschman L. Translating biology: text mining tools that work. Pac Symp Biocomput 2008;13:551–5.

Guest Editors

Wendy W. Chapman

Department of Biomedical Informatics,
University of Pittsburgh, 200 Meyran Avenue,
Pittsburgh, PA 15260, USA
Fax: +1 412 647 7190.
E-mail addresses: wec6@pitt.edu, wendy.w.chapman@gmail.com

K. Bretonnel Cohen

Center for Computational Pharmacology,
Biomedical Text Mining Group,
University of Colorado School of Medicine, Denver, CO 80202, USA