RC23040 (W0312-100) December 17, 2003
Computer Science

IBM Research Report

SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics

Eric W. Brown
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

William Hersh
Department of Medical Informatics & Clinical Epidemiology
OHSU BICC
3181 SW Sam Jackson Park Road
Portland, OR 97239

Alfonso Valencia
Protein Design Group
C.N.B. - C.S.I.C.
Campus Universidad Autonoma
Cantoblanco 28049 Madrid
Spain

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

SIGIR 2003 Workshop on
Text Analysis and Search for Bioinformatics

Eric W. Brown
IBM TJ Watson Research Center
PO Box 704
Yorktown Heights, NY 10598
ewb@us.ibm.com

William Hersh
Dept. of Medical Informatics & Clinical Epidemiology
OHSU BICC
3181 SW Sam Jackson Park Rd.
Portland, OR 97239
hersh@ohsu.edu

Alfonso Valencia
Protein Design Group.
C.N.B. - C.S.I.C.
Campus Universidad Autonoma
Cantoblanco 28049 Madrid
valencia@cnb.uam.es

Introduction

Bioinformatics is generally defined as the application of information technology to help solve problems in cellular and molecular biology. This covers a broad range of topics from computational models of protein folding to the storage, search, and retrieval of gene sequence data. An emerging topic of interest in this area is automatic analysis of the bio-medical scientific literature. The goals in this area generally include providing easy access to specific textual information from a potentially very large corpus, and automatically extracting information from the text in a form amenable to further, possibly more structured analysis.

The bio-medical literature is full of papers that describe clinical and experimental results, many of which are expressed at the cellular or molecular level as interactions between genes, proteins, and other molecules, or as signal pathways through the cell. Scientists typically describe these results using complex natural language. If this information can be accurately extracted and represented in a more structured form, it can be used to facilitate locating the source document, and, perhaps more interestingly, it can form the basis of a richer knowledge representation and analysis system.

Techniques developed for the scientific literature may also be applicable to the Medical Informatics domain, which includes clinical patient records. Clinical records contain the observations of clinicians as well as the results of medical tests. This may include coded or structured information, but important details often reside in textual notes. Applying text analysis and information extraction techniques can help automate tasks currently performed manually, enable various statistical analyses on individual and large groups of records, and allow connections back to the bioinformatics world. This last task will become more important as personalized medicine (e.g., individually customized drugs) evolves.

Over the last several years interest in the application of text analysis and natural language processing techniques to bio-medical text has grown rapidly, and a research community of bio-medical scientists, computer scientists, and computational linguists has emerged. In light of this trend, we proposed this workshop with two goals in mind. First, we wanted to provide a forum where the latest problems, techniques, and results in Bioinformatics for text can be discussed. Second, we wanted to bring together the Bioinformatics and SIGIR communities to share their insights and results and build on each other's work.

We are pleased to report that both goals were met. The workshop attracted over thirty participants, including researchers in information retrieval, text analysis, bioinformatics, and even a few molecular biologists. We selected nine papers for presentation at the workshop, and scheduled plenty of time for questions and discussion. Below we summarize the papers and conclude with the main themes and issues that were discussed.

Presentations

After a brief introduction by Eric Brown (IBM Research), Rohini Srihari (SUNY at Buffalo) presented the first paper, "Concept Chain Graphs: A Hybrid IR Framework for Biomedical Text Mining." Srihari et al. describe a system that combines traditional information retrieval with information extraction in a new framework called concept chain graphs (CCGs). The CCG is a probabilistic network of automatically identified concepts linked by automatically extracted relationships. It provides a rich representation of a document collection and facilitates exploration and visualization of the concepts described within those documents. The CCG provides the foundation for a number of applications that allow end users to perform *unapparent information revelation* (UIR). The key notion is that after performing a traditional ad-hoc search, a bio-medical researcher can benefit substantially from the ability to filter and browse results and, in particular, visualize and explore attributes of concepts and relationships between concepts in the domain. Srihari et al. present experimental results showing how their technique can be used to explore gene attributes and to discover new associations between genes through these attributes.

The next presentation, "An Evaluation of Unnamed Relations Computation for Discovery of Protein-Protein Interactions," was prepared by James Cooper (IBM Research). Cooper could not attend the workshop, so he submitted his presentation as a narrated slide show and was available via chat to answer questions. Cooper's paper describes work on automatically extracting protein-protein interactions between yeast proteins. The work involves identifying mentions of proteins in the text and exploring a number of ways to identify and extract descriptions of interactions between these proteins. Cooper performs protein mention identification with a dictionary of protein names and synonyms. The methods for identifying interactions include statistical co-occurrence and syntactic analysis of noun-verb-noun constructs. To evaluate these techniques, Cooper created a test-bed of 564 protein interactions, derived from the Munich Information Center for Protein Sequences[1]. In his preliminary results, Cooper found that syntactic analysis did not improve the results obtained with statistical co-occurrence techniques, and that optimal performance was achieved when the co-occurrence window was two sentences. Cooper also presented a graphical visualization of protein-protein interactions that allowed interactive exploration of these relationships. Using this visualization, a user can discover meaningful secondary relationships between proteins.

William Hersh (Oregon Health & Science University) gave the next presentation, "Of Mice and Men (and Rats and Fruit Flies): The TREC Genomics Track." Hersh is the chair and primary organizer of the new Genomics track[2] at TREC 2003, and his presentation provided an overview of that track. Sponsored by the National Institute of Standards and Technology, TREC[3] is an annual conference for the evaluation and discussion of various techniques and tasks in the broad area of information retrieval. Given the growing activity and interest in the area of text analysis for bio-medical data, Hersh and others recognized the need for a forum for systematically evaluating techniques in this domain. In its inaugural form, the Genomics track comprised two tasks. The primary task was similar to the traditional ad-hoc search task: given a gene, find all documents in a collection of MEDLINE abstracts that discuss a function of that gene. The secondary task was more of an information extraction or summarization task: given a gene and a document known to describe the gene's function, automatically extract the function description. The definition of these two tasks for the first year of the track was driven by a need to define tasks that were meaningful in the domain, yet could be evaluated in an automatic fashion. Automatic evaluation implies relevance judgements for the task, and these judgements were derived from the Gene Reference Into Function (GRIF) entries from LocusLink[4], a database of gene information publicly available and maintained by the

[1] http://mips.gsf.de

[2] http://medir.ohsu.edu/~genomics/

[3] Text REtrieval Conference, http://trec.nist.gov

[4] http://www.ncbi.nlm.nih.gov/LocusLink/

National Center for Biotechnology Information[5]. A GRIF is an entry in the LocusLink record for a gene that includes a citation to a MEDLINE abstract that describes some function of the gene, along with a textual summary of that function. Thus, the GRIF entries serve to both define the set of relevant documents for the primary task, and provide the reference text for the secondary task. In addition to describing these tasks in detail, Hersh described the five-year plan for the track and reported that 56 groups had registered for the inaugural Genomics track.

Padmini Srinivasan (University of Iowa) was to present "Mining MEDLINE Metadata to Explore Genes and their Connections," but was unable to attend the workshop at the last minute. In their paper, Srinivasan et al. present a method for discovering and exploring connections between genes using metadata from MEDLINE abstracts that discuss the genes. Srinivasan et al. have implemented their method in a system that comprises three phases. In the first phase, the end user specifies a dataset for analysis. A dataset is a collection of topics to be analyzed by the system, and each topic is formed by a search specification and its corresponding set of retrieved documents (e.g., a PubMed[6] query and the resulting MEDLINE abstracts). In the second phase, the system extracts MeSH[7] terms (the metadata) from each document in the topic set, groups the terms according to semantic type (as defined by UMLS[8]), and creates a profile vector for each semantic type. A profile vector consists of term weights for each MeSH term in the semantic type. The system uses the profile vectors to compute a number of statistics for the dataset, including topic distance, topic co-occurrence, and profile similarity. These statistics are available during phase three of the analysis, where the user can query and visualize the statistics in a variety of ways. For example, Srinivasan et al. show how they have used the system to automatically discover the relationship between Raynaud's disease and fish oil, which was first discovered in Swanson's work on hidden links (Swanson 1986).

Moving out of the technical literature domain and into the domain of medical records, Ilya Goldin (University of Pittsburgh) presented "Learning to Detect Negation with 'Not' in Medical Texts," which explores the challenging problem of automatically understanding negation in natural language text. In particular, Goldin addressed the problem of determining whether or not an occurrence of the word 'not' negates the meaning of a recognized UMLS term that appears nearby. Goldin et al. described a baseline system, NegEx (Chapman, Bridewell et al. 2001), that always negates a UMLS term when it co-occurs with 'not' in a window of six words or phrases. The challenge was to improve the precision of this baseline system by exploring machine learning techniques, including Naïve Bayes and Decision Trees, to determine when 'not' correctly predicts negation. The conclusion was that the both machine learning techniques provided a statistically significant improvement in accuracy over the baseline. The Decision Tree approach was particularly interesting in that it revealed the following simple rule to improve the baseline: when negation of a UMLS term is triggered with the negation phrase 'not,' if the term is preceded by 'the' then do not negate.

Debra Burhans (Canisius College) and Alistair Campbell (Hamilton College) presented "Exploring the Role of Knowledge Representation and Reasoning in Biomedical Text Understanding," which details an ambitious goal of building a knowledge representation and reasoning system that can automatically infer new knowledge from a small, focused set of abstracts. Burhans and Campbell are collaborating with Gary Skuse (Rochester Institute of Technology), an expert on the disease neurofibromatosis type 1 (NF1). Thus, they have chosen NF1 as their initial domain. At this stage of their work, Burhans et al. are primarily concerned with validating the efficacy of their overall approach. They have selected the SNePS

---

[5] http://www.ncbi.nlm.nih.gov/

[6] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

[7] http://www.nlm.nih.gov/mesh/meshhome.html

[8] http://www.nlm.nih.gov/research/umls/

(Shapiro and Rapaport 1992) knowledge representation and reasoning system for their knowledge base, and are manually transcribing abstracts about NF1 and loading them into SNePS. Once the knowledge base is built up from these abstracts, the authors can explore the accuracy and effectiveness of the overall system, exploiting Skuse's domain expertise to validate their results. Should this approach prove viable, they recognize that hand-transcribing abstracts into the knowledge base will not scale, so the next step will be to develop and explore techniques to automatically analyze, transcribe, and load abstracts into the knowledge base.

Yoshiaki Kawasaki (University of Tokyo) presented "Extracting Biomedical Ontology from Textbooks and Article Abstracts." This work addresses the problem of how analyze text and automatically construct ontologies of bio-medical domain concepts. While a number of researchers have tried to automatically extract ontologies from text, the resulting ontologies are often disconnected and lacking in structure. The key idea put forth by Kawasaki et al. is to start with textbooks, where the material is typically presented in an orderly fashion from general to more specific topics. By recognizing this progression from general to specific topics, the extracted concepts can be organized in a more structured, hierarchical ontology. Kawasaki et al. presented results comparing ontologies built from textbooks only, research papers only, and a combination of the two. They also attempted to compare the automatically built ontologies with the well-known Gene Ontology[9] (GO), using GO as the "gold standard." Starting with textbooks yields ontologies with greater average depth than analyzing abstracts alone, suggesting that Kawasaki et al.'s approach does address the problem of disconnected and unstructured ontologies. They were, however, unable to map the nodes in their automatically generated ontology into GO and perform a meaningful comparison. Thus, their results are preliminary and suggestive at best. Nevertheless, the technique is interesting, and the results underscore a common theme at the workshop, which is a need for better evaluation methods and standard test sets.

Continuing on the topic of exploiting GO, Cornelia Verspoor (Los Alamos National Lab) presented "The Gene Ontology as a Source of Lexical Semantic Knowledge for a Biological Natural Language Processing Application." Verspoor et al. explore the feasibility of using GO as a lexical and semantic resource for natural language processing in the domain, i.e., does GO provide sufficient coverage at the lexical level to support recognizing terms in the domain and, more importantly, can the semantics inherent in the ontology be exploited. To answer these questions, Verspoor created a corpus of approximately 10,000 MEDLINE abstracts and compared the overlap of terms in the corpus with terms in GO. The overlap at the token level, even without stemming, was quite good for the high and middle frequency terms in the corpus. To take advantage of the semantic information in GO (e.g., *isa* and *part of* relationships, associations between terms and gene products, etc.), the tokens in the corpus must map correctly to nodes in the GO hierarchy. The nodes in the GO hierarchy are described by multi-word phrases, but the overlap at the phrase level between GO and the corpus was relatively small. Verspoor et al., therefore, propose two styles of mapping the text into GO: *direct* and *indirect*. Direct mappings occur when there is an exact match at the phrasal level. To make indirect mappings, Verspoor et al. describe a number of methods for transforming the explicit GO relationships and deriving semantic relationships between shorter phrases or single words. The next step in their work is to use these direct and indirect mappings in their natural language processing system and evaluate the effectiveness of this approach for incorporating lexical and semantic information.

In the final presentation of the workshop, Zhonghua Yu (University of Tokyo) presented "Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense Per Discourse Hypothesis." Yu et al. address the problem of disambiguating abbreviations in the domain by combining a Support Vector Machine (SVM) classifier with the "One Sense Per Discourse" hypothesis (Gale, Church et al. 1992), which states that the vast majority of occurrences of an ambiguous

[9] http://www.geneontology.org/

term in a single discourse will have the same meaning. Applied to this domain, the hypothesis suggests that all occurrences of an ambiguous abbreviation in a single MEDLINE abstract will have the same meaning, or long form. Using this hypothesis, one can automatically generate training and test data by selecting a set of abbreviations and, for each abbreviation, finding abstracts that contain both the abbreviation and the long form in a definitional context. One can then assume that all other occurrences of the abbreviation in the abstract map to the definitional long form. After training an SVM classifier for each abbreviation, it is still possible that the classifier will incorrectly classify multiple occurrences of the same abbreviation in a single abstract into multiple long forms. Yu et al., therefore, apply the One Sense Per Discourse hypothesis again at run time and incorporate a voting component into the classification methodology. The voting component improves abbreviation disambiguation accuracy by 2% over the baseline SVM classifier without voting.

Conclusions

The variety of presentations, large attendance, and enthusiasm of the workshop participants confirmed that text analysis for the bio-medical domain is an active, important area of research. The presentations, in particular, show that we must apply sophisticated information retrieval and natural language processing techniques to the problems in this domain if we are to advance the state-of-the-art in the area of searching, analyzing, summarizing, and extracting information from bio-medical text. Moreover, we must apply these techniques in combination with domain expertise to ensure that we are solving the right problems and generating useful solutions.

Overall, four main themes rose out of the workshop. First, there is an urgent need for standard evaluation test-beds in this area. While some progress has been made in this area (e.g., the GENIA corpus[10], the 2002 KDD Cup[11], the TREC Genomics track), researchers continue to build ad-hoc test collections and report results that are difficult to replicate or compare. Test collections must clearly define important and relevant tasks that advance the state-of-the-art, they must provide the correct answers (relevance judgements or a "gold standard") for automatic evaluation, they must describe meaningful evaluation metrics, and they must be widely available and easily obtainable. Test collections should also address problems at a variety of levels, from low-level methods (e.g., named entity identification) that will be common to many complex systems, to high-level tasks that address typical end-to-end tasks performed by end users of the systems being evaluated.

Second, while functionality and effectiveness are critical, we cannot ignore scalability and performance. If we are truly going to build technologies that will help bio-medical researchers solve important problems in their domain, from the start these technologies must scale to huge volumes of text and provide high throughput and fast response times. We also need to be mindful of all the different data sources that must be processed – MEDLINE is not the only source of bio-medical text. Currently MEDLINE contains over 12 million citations and the abstracts include approximately 40GB of XML text. These are just the abstracts; the corresponding full text articles would be much larger. Other sources of bio-medical text include patents and patient records. Some healthcare institutions have terabytes of text in patient medical records. While these figures may pale in comparison to the scale of the Web (which we can easily search in under a second with Google[12]), the challenge is to go beyond text indexing and apply computationally intensive natural language processing and text analysis to this data.

The third theme emphasized in the workshop is the need for domain expertise in both text analysis and the bio-medical domain to ensure that we are solving the right problems and producing useful solutions. In

[10] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

[11] http://www.biostat.wisc.edu/~craven/kddcup/index.html

[12] http://www.google.com/

particular, bio-medical expertise and advice is critical for any project that will make a meaningful contribution to this area. This is even more important for systems that attempt to extract structured information from text and then apply additional reasoning to either use the information as a knowledge reference or apply reasoning over that knowledge to infer new, previously unknown concepts and relationships. Domain expertise is required to validate these approaches and confirm the results.

Finally, the fourth theme was that there is clearly a continuing need for forums such as this workshop where practitioners from information retrieval / natural language processing and the bio-medical / Bioinformatics domains can gather and share their expertise. Both communities already have a significant body of prior art. Sharing and building on top of this history will ultimately produce the most rapid advances and effective solutions.

Acknowledgements

We wish to thank the SIGIR conference organizers for supporting this workshop and running an excellent conference, even in the face of bio-medical adversity. We would also like to thank the workshop participants and especially our presenters for creating an interesting and stimulating program.

References

Chapman, W. W., W. Bridewell, et al. (2001). "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries." Journal of Biomedical Informatics **34**(5): 301-310.

Gale, W. A., K. W. Church, et al. (1992). One Sense Per Discourse. ARPA Workshop on Speech and Natural Language Processing.

Shapiro, S. C. and W. J. Rapaport (1992). "The SNePS Family." Computers and Mathematics with Applications **23**: 243-275.

Swanson, D. R. (1986). "Fish-oil, Raynaud's Syndrome, and Undiscovered Public Knowledge." Perspectives in Biology and Medicine **30**(1): 7-18.