Dai HJ, Chang YC, Tsai RTH *et al.* New challenges for biological text-mining in the next decade. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 169–inside back cover Jan. 2010

New Challenges for Biological Text-Mining in the Next Decade

Hong-Jie Dai[1,2], Yen-Ching Chang[1], Richard Tzong-Han Tsai[3], and Wen-Lian Hsu[1,2], *Fellow, IEEE*

[1] *Institute of Information Science, "Academia Sinica", 115, Taiwan, China*

[2] *Department of Computer Science, "National Tsing-Hua University", 300, Taiwan, China*

[3] *Department of Computer Science and Engineering, Yuan Ze University, 320, Taiwan, China*

E-mail: {hongjie, ro3789, hsu}@iis.sinica.edu.tw; thtsai@saturn.yzu.edu.tw

Received September 1, 2009; revised November 24, 2009.

Abstract The massive flow of scholarly publications from traditional paper journals to online outlets has benefited biologists because of its ease to access. However, due to the sheer volume of available biological literature, researchers are finding it increasingly difficult to locate needed information. As a result, recent biology contests, notably JNLPBA and BioCreAtIvE, have focused on evaluating various methods in which the literature may be navigated. Among these methods, text-mining technology has shown the most promise. With recent advances in text-mining technology and the fact that publishers are now making the full texts of articles available in XML format, TMSs can be adapted to accelerate literature curation, maintain the integrity of information, and ensure proper linkage of data to other resources. Even so, several new challenges have emerged in relation to full text analysis, life-science terminology, complex relation extraction, and information fusion. These challenges must be overcome in order for text-mining to be more effective. In this paper, we identify the challenges, discuss how they might be overcome, and consider the resources that may be helpful in achieving that goal.

Keywords bioinformatics database, mining method and algorithm, text mining

1 Introduction

Life-science journal publishing has undergone a digital revolution in the last decade. The massive flow of scholarly publications from traditional paper journals to online outlets has benefited biologists in the ease of access, but has also left these scholars adrift in the deluge of biological literature they have made available. Recent biology contests, such as JNLPBA[1] and BioCreAtIvE[2-3], have evaluated ways in which the literature may be navigated. Among the methods evaluated, text mining has shown the most promise because it makes biological literature more accessible, and therefore more useful[4-5].

Text mining involves analyzing a large collection of documents in a manner that reveals specific information, such as the relationships and patterns buried in the collection, which is normally imperceptible to readers. A key text mining task involves linking extracted information to form new facts or new hypotheses that can be explored further by more conventional means of experimentation[6].

In the biomedical domain, several tools[7-9], competitions[10-12] and projects[13-14] have started to incorporate text mining technology. However, text mining is difficult to implement in many cases because the vital components of scientific communication — journals and databases — are designed to be read by people, not computers. Computers cannot extract information efficiently from unstructured text, which is the format adopted by most journals and databases. Fortunately, some publishers, e.g., the Public Library of Science (PLoS) and BioMed Central, have sought to address this problem by making the full texts of their publications available as downloadable XML files that can be processed easily by computer programs. The FEBS Letters journal is currently experimenting with embedding text-mining systems (TMSs) in the manuscript submission process to construct structured digital abstracts semi-automatically[15] — machine-readable XML summaries of pertinent facts in the published articles.

With the recent advances in text-mining technology and the fact that some publishers are now making the full texts of articles available in XML format, TMSs

Regular Paper

This work was supported by the "National Science Council" under Grant Nos. NSC 97-2218-E-155-001 and NSC96-2752-E-001-001-PAE, the Research Center for Humanities and Social Sciences, and the Thematic Program of "Academia Sinica" under Grant No. AS95ASIA02.

can be applied to the full texts rather than just the abstracts, and to accelerate literature curation, maintain the integrity of information, and ensure proper linkage of data to relevant resources. However, several new challenges have emerged in relation to full text analysis, life-science terminology, complex relation extraction, and information mergence. First, terms, such as gene names and corresponding database identifiers, are so numerous and varied that even specialists have difficulty understanding them and keeping track of updates and revisions. While state-of-the-art normalization systems developed for BioCreAtIvE II[16] can normalize gene identifiers for humans relatively well, such systems have yet to be developed for inter-species normalization. Second, full-text analysis requires the use of more sophisticated natural language processing (NLP) techniques than current biological information retrieval and extraction tools can handle[17-18]. For example, in full text analysis, TMSs must extract cross-sentence relations, while most current TMSs can only extract relations within sentences. Third, current TMSs are unable to merge information from disparate sources with different contextual and typographical representations. However, associating works in the literature via pathway information is essential. In the next section, we discuss the above-mentioned challenges in more detail.

2 Full Text Processing

Most Biological Natural Language Processing systems have only been applied to abstracts because of the latter's availability and abridged nature. Abstracts are good targets for information extraction (IE) as they summarize the content of articles. However, the full texts of papers contain more information, relevant or not, which should be treated carefully[19].

The preliminary results of applying current state-of-the-art TMSs to full texts showed a promising F-score[①] of 28.85%[20] in the BioCreAtIvE (critical assessment for information extraction in biology) protein-protein interaction (PPI) annotation extraction task[21]. However, they also revealed several issues of concern:

1) Errors resulting from converting PDF or HTML formatted documents to plain text.

2) Difficulties in processing tables and figure legends.

3) Multiple references to organisms and the resulting inter-species ambiguity in gene/protein normalization.

4) Sentence boundary detection errors.

5) Difficulties in extracting the associations and handling the coordination of multiple interaction pairs in single sentences.

6) Phrases used to describe interactions in legends or titles that do not correspond to grammatically correct sentences in the text.

7) Errors in shallow parsing and POS (part-of-speech)-tagging tools trained on general English text collections when applied to specific expressions and abbreviations found in biomedical texts.

The open text mining interface, a project directed by the Nature Publishing Group, helps solve the data format conversion errors and difficulties mentioned in issues 1) and 2) because it provides open access to full text documents published in XML format. The full-text versions of scientific literature that are machine-readable, but many other aspects need to be improved further, as we explain in the following subsections.

2.1 Named Entity Identification in Full Text

2.1.1 Named Entity Recognition

The fundamental task of recognizing biological terms, such as gene and protein names, is the first step towards making full use of the information encoded in biomedical texts. The named entity recognition (NER) task in the biomedical domain has different characteristics from that in the newswire domain, such as the MUC-7 NER task[22]. The unique difficulties of biomedical NER are as follows. First, the number of new gene names is growing continually, and it is hard to recognize all of them because there is so much inconsistency among them[20]. Second, authors do not use standardized names; they prefer to use abbreviations or other forms depending on personal inclination[23]. Because of their limited length, abbreviations/acronyms are often identical to the respective genes' symbols and thus increase the ambiguity of the nomenclature[24]. For instance, 80% of the abbreviations listed in the UMLS have ambiguous versions in MEDLINE[25]. Third, gene names are similar to/occur with other terminology varying from gene/protein names, such as the names of cells, tissues or organs[26]. For example, C1R is a cell line, but it is also a gene (SwissProt P00736). TMSs must be able to distinguish between different genes with identical names as well as to determine whether certain gene names refer to completely different biological entities like viruses. For compound names, it is also necessary to determine where the name begins and ends within a sentence. The task can be particularly difficult when verbs and adjectives are embedded in names[27].

A large number of machine learning algorithms have been developed to deal with the NER problems; for example, the hidden Markov model[28], the support vector machine model[29], the maximum entropy Markov model[30] and the conditional random

[①] F-score is the weighted harmonic mean of precision and recall.

field model[31]. To capture the diverse characteristics of biomedical entities, several feature sets, including lexicons, orthographic/affix information, and even external resources like the WWW have been incorporated into different algorithms. It is conceivable that the recognition results derived by these algorithms will be diverse but complementary to each other. One natural idea for improving the performance of biomedical NER is to combine the results of several algorithms. The results of the BioCreAtIvE II gene mention task[20] and those reported by Si *et al.*[32] show that it is possible to achieve higher recognition accuracy by combining the results of multiple NER algorithms.

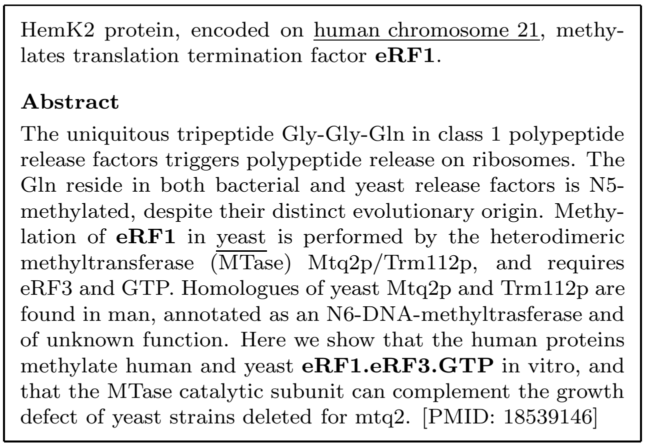
False positive gene/protein names found in the full texts of articles pose great challenges for TMSs in such basic tasks as identifying gene and protein names in biomedical texts. Broadening the range of entities beyond genes/proteins to include entities like chemicals and diseases[33-34] can resolve the problem. Identifying these entities also allows us to consider biologically relevant relations, such as which entities they are derived from, where they are located, which have agency in which processes, or which participate in what processes.

In addition, it may also be possible to use algorithms that can identify acronyms/abbreviations to extract acronyms from text automatically without checking whether they overlap with the gene nomenclature. Although several algorithms have been proposed for this purpose[35-36], only a few can extract acronyms and disambiguate gene names[37]. We hope that integrating these tools will improve the NER performance.

2.1.2 Inter-Species Normalization

Gene normalization (GN) determines the unique identifiers of genes and proteins mentioned in the literature. The concept was inspired by a step in a typical curation pipeline for model organism databases. After an article has been selected for curation, curators list the genes or proteins of interest in this article[16]. Although the concept of GN was inspired by curation, the BioCreAtIvE I/II computer-aided GN task[16,38] oversimplified curation and performed GN by normalizing genes in abstracts rather than on the full-text. Actually, human curators normally work on the full texts and only identify particular kinds of genes of interest. Cohen *et al.*[39] proposed a computer-aided GN system that, given a document, provides a ranked list of genes that are discussed in the document. The BioCreAtIvE II.5 competition of 2009[40] included a similar ranking task. Such a ranked list could be used as an aid by human curators.

Computer-aided GN presents several difficult problems that need to be solved in order to reduce the workload of human curators. First, gene and protein names often have several spelling variations or abbreviations. Second, gene products are often described indirectly via phrases, such as "light chain-3 of microtubule-associated proteins 1A and 1B", instead of by specific names or codes. A number of approaches[41-42] have been proposed to address these problems in the BioCreAtIvE I/II's GN task. The evaluation results provide some insight into how these problems affect our capacity to normalize the genes mentioned in biological abstracts, but GN is not yet practical. The BioCreAtIvE I/II GN task involved normalizing various abstracts and demonstrating how much TMSs' success varied according to the organism discussed in the abstracts. The results showed that the performances were satisfactory for normalizing abstracts that mentioned the genes and proteins of humans (F-score 0.81), mice (F-score 0.79), yeast (F-score 0.92) and flies (F-score 0.82), respectively. However, the task did not address the important issue of inter-species GN, which exists in many published articles.



HemK2 protein, encoded on human chromosome 21, methylates translation termination factor **eRF1**.

Abstract

The uniqitous tripeptide Gly-Gly-Gln in class 1 polypeptide release factors triggers polypeptide release on ribosomes. The Gln reside in both bacterial and yeast release factors is N5-methylated, despite their distinct evolutionary origin. Methylation of **eRF1** in yeast is performed by the heterodimeric methyltransferase (MTase) Mtq2p/Trm112p, and requires eRF3 and GTP. Homologues of yeast Mtq2p and Trm112p are found in man, annotated as an N6-DNA-methyltrasferase and of unknown function. Here we show that the human proteins methylate human and yeast **eRF1.eRF3.GTP** in vitro, and that the MTase catalytic subunit can complement the growth defect of yeast strains deleted for mtq2. [PMID: 18539146]

Fig.1. Abstract (PMID 18539146) in PubMed.

The extract in Fig.1, taken from an abstract in PubMed, exemplifies the challenges posed by many articles when using computer-aided GN. One name, abbreviation or code, may refer to genes in multiple species, each with its own unique ID, or even to multiple genes in the same species or across different species. For example, the abstract (PMID: 18539146) in Fig.1 discusses the methylation of the gene "eRF1". In the UniProt database, the gene's name is listed as a synonym of multiple genes, such as ZFP36L1 (SwissProt Q07352) and ETF1 (SwissProt P62495) even though their functions are different. Moreover, both ZFP36L1 and ETF1 refer to multiple species, namely humans, mice, and rats. We also observe that "eRF1" appears in the title as a human gene and in the third sentence of the abstract as a yeast gene. Finally, the complex "eRF1.eRF3.GTP" in the last sentence is a protein complex and should not be associated with any database identifiers. These few sentences illustrate how much more GN needs to be improved before it can be used in practice.

2.2 Relation and Fact Extraction from Full Texts

TMSs, like human curators, should work on full text articles. The information provided in the headings, figure legends, and tables of full text articles helps TMSs extract relations and facts; and may help users discover implicit associations between genes and diseases in the future[18]. Seki and Javed[43] conducted a small preliminary experiment and reported that using the full text articles, rather than just their abstracts, to extract gene-disease relations greatly improved the ability of their text-mining system to discover facts and relations. In addition, Cooper and Kershenbaum[44] conducted a detailed study of 65 abstracts and found that some PPIs were only reported in the full texts of the respective papers. The abstracts of some papers did not contain any protein names. Hence, TMSs should analyze the full text articles, not just their abstracts. In the following sections, we consider the challenges that must be addressed.

2.2.1 Relevant Versus Irrelevant Information

TMSs need to distinguish between relevant and irrelevant information, but different criteria may have to be applied depending on which section of an article the text-mining system is analyzing or mining. Shah *et al.*[45] demonstrated that there are substantial differences in the content of different sections of a publication. For example, specific terms, like the names of certain genes, may be mentioned in the titles of articles in a paper's bibliography, but TMSs should disregard such terms. To identify useful terms, TMSs should compare terms mentioned in papers' abstracts, which usually contain a high density of relevant terms (keywords), to terms appearing throughout the full texts of the respective papers.

Moreover, TMSs should also be able to associate useful pieces of information in the legends of figures and tables with the text of the article, but this task is quite challenging. One reason is that figures and images often have multiple sub-figures, so TMSs must be able to identify the sub-figures and match each one with the appropriate sentences or references in the text. Although this task may seem difficult, TMSs that have such a capacity might discover more useful relations or facts than those normally extracted. Some researchers have been successful in combining text-mining and image recognition techniques[46-47]; however, there is a need for much greater collaboration between researchers in the two fields before TMSs can perform image recognition and mine related text easily.

2.2.2 Relation Extraction

In the biomedical field, researchers are interested in PPIs, gene-gene interactions and protein-disease interactions. The major goal of relation extraction is to discover the relations embedded within sentences, paragraphs, or entire documents. Currently, the most popular relation extraction approaches include rule-based[48-49], kernel-based[50-51], and co-occurrence-based[52-53] methods. Most works focus on identifying the relations between proteins[53-55]. Craven and Kumlien[56] identified the relations between proteins and sub-cellular locations; while Rindflesch *et al.*[57] extracted the relations between cancer-related genes, drugs and cell lines. Less work has been done on extracting the relations between genes and diseases[58-59], but the area is now attracting more research efforts.

Among existing methods, employing parsers to analyze syntactic and semantic structures is useful. Miyao *et al.*[60] performed a comparative evaluation of state-of-the-art syntactic parsing methods, including dependency parsing, phrase structure parsing and deep parsing, and their contribution to PPI extraction. The study provides researchers with a good reference for choosing appropriate parsers for their work. However, there is no guarantee that the results reported by Miyao *et al.* can be generalized to other datasets and tasks.

The results of the BioCreAtIvE II PPI task[21] demonstrate that current TMSs can detect binary relations in abstracts reasonably well[49,61], but they are not always as effective in extracting significant relations from full-text articles. There are three reasons for this phenomenon.

First, biomedical terms, such as gene names, may have different meanings in full texts depending on the context or the section in which they appear. The same gene in one section may belong to different species (consider the example shown in Fig.1). Second, the frequent use of synonyms, abbreviations, and acronyms in biomedical texts hinders semantic analysis. For instance, extracting facts from the Results section may require resolving acronyms or synonyms only mentioned in the Introduction section. Third, biomedical texts usually contain several compound nouns as well as noun phrases linked by prepositions. Fourth, TMSs have difficulty when one or more proteins involved in an interaction are expressed by more than one sentence; or when they are expressed using anaphora, as shown in the following example:

Human growth hormone (**hGH**) binds to _its_ receptor (**hGHr**) in a three-body interaction: one molecule of _it_ and two identical monomers of the receptor from a trimer.

Many papers have addressed relation extraction, summarization, and evaluation issues, but few have focused on co-reference (anaphora) resolution[62], possibly because there are few publicly available datasets for system building and evaluation. Despite the substantial amount of annotation work carried out on co-referencing in molecular biology, few biomedical corpora with co-reference annotations are currently available[63]. Recently, the GENIA corpus was annotated with co-references. Nguyen *et al.*[64] conducted a pioneering study of the differences between newswire and biomedical co-reference annotated corpora. We look forward to the integration of more sophisticated NLP techniques in this respect.

3 Future of Text-Mining Applications

3.1 User-Focused Applications

Text-mining researchers are typically good at analyzing textual content, but they are not as good at building interactive systems that users can adopt easily[33]. To resolve the problem, researchers must design applications with intuitive interfaces that require little or no knowledge of text-mining and NLP technology. The objective is to provide bioinformatics, biological, biomedical, and pharmacological researchers with a high-level view of biological interactions and help them form new hypotheses. The useful PubMed-EX browser extension[65], shown in Fig.2, is an example of such an effort.

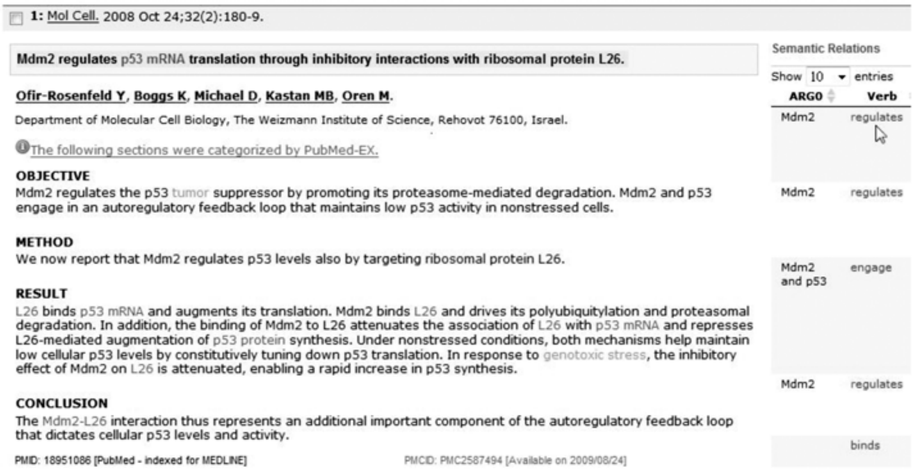
PubMed-EX annotates onsite PubMed search results with additional text-mining information but users do not pay any extra effort such as to learn how to input a specific query. Currently, its processing speed is quite slow, but it does hide the complicated text-mining technology on which it is based.

Text-mining researchers should strike a compromise between the accuracy of text-mining results and the overall processing speed. Obviously, full text analysis requires more computational capacity and time than the analysis of abstracts. Users may accept a processing time of 10 minutes per article for off-line processes, such as database curation, but they may not be as patient when it comes to on-line services that provide semantic annotations or relation extraction. Therefore, providing on-the-fly full text processing, while maintaining a satisfactory accuracy level, remains a challenge for text-mining researchers.

Certain types of users, such as content providers and corpus annotators, require interfaces that allow them to change annotations, dredge for information, link resources, and create new information resources to capture new concepts[33]. The research community requires more collaborative annotation and up-to-date knowledge in biological databases, but it does not have the tools that make these procedures easy to implement. We discuss this issue in the next subsection.

3.2 Integration, Communication and Collaboration

Bioinformatics researchers often need to consult numerous databases and web servers, but many find integrating heterogeneous datasets from disparate databases associated with multiple web servers a daunting task[66]. To integrate biological data from multiple



□ **1:** Mol Cell. 2008 Oct 24;32(2):180-9.

Mdm2 regulates p53 mRNA translation through inhibitory interactions with ribosomal protein L26.

Ofir-Rosenfeld Y, Boggs K, Michael D, Kastan MB, Oren M.

Department of Molecular Cell Biology, The Weizmann Institute of Science, Rehovot 76100, Israel.

The following sections were categorized by PubMed-EX.

OBJECTIVE

Mdm2 regulates the p53 tumor suppressor by promoting its proteasome-mediated degradation. Mdm2 and p53 engage in an autoregulatory feedback loop that maintains low p53 activity in nonstressed cells.

METHOD

We now report that Mdm2 regulates p53 levels also by targeting ribosomal protein L26.

RESULT

L26 binds p53 mRNA and augments its translation. Mdm2 binds L26 and drives its polyubiquitylation and proteasomal degradation. In addition, the binding of Mdm2 to L26 attenuates the association of L26 with p53 mRNA and represses L26-mediated augmentation of p53 protein synthesis. Under nonstressed conditions, both mechanisms help maintain low cellular p53 levels by constitutively tuning down p53 translation. In response to genotoxic stress, the inhibitory effect of Mdm2 on L26 is attenuated, enabling a rapid increase in p53 synthesis.

CONCLUSION

The Mdm2-L26 interaction thus represents an additional important component of the autoregulatory feedback loop that dictates cellular p53 levels and activity.

PMID: 18951086 [PubMed - indexed for MEDLINE] PMCID: PMC2587494 [Available on 2009/08/24]

Semantic Relations

Show 10 entries

ARG0	Verb
Mdm2	regulates
Mdm2	regulates
Mdm2 and p53	engage
Mdm2	regulates
	binds

Fig.2. A PubMed abstract annotated with text-mining results by PubMed-EX.

heterogeneous databases, researchers have adopted two major approaches: centralization[67] and decentralization[68]. However, the integration efforts have been piecemeal and have only considered a fraction of bioinformatics data, so complex queries remain challenging. Integrating data from multiple databases and analyzing it via TMSs is difficult. Zhang *et al.*[66] proposed a Web 2.0[69] based model that represents a shift in focus from working locally to working in networked settings. Under this new approach, the Web is seen as a social, collaborative, and collective space. The model provides a vision of the future, where annotation will be performed collaboratively and innovative web tools will support such collaboration. Further development of tools like WikiProtein[70] and CBioC[71], which support collaborative annotation is essential.

3.3 Information Fusion

With the advent of advanced TMSs, researchers may be able to integrate mined information and thereby gain more insight into biological literature. The most critical biological reactions are recorded in "pathways," which include a myriad of cellular or disease events with multiple protein-protein relationships and tend to influence each other directly. However, for a number of reasons, TMSs have trouble fusing mined information to reveal pathways.

First, mapping named entities to nodes in pathways requires highly context dependent properties. Named entities (NEs) may have different meanings in the same context. For example, an NE may be located in the nucleus, in the cytoplasm, or on the cell membrane. It may also refer to a cellular function, in which case it might be phosphorylated or acetylated. Thus, two consecutive sentences may mention a named entity, but the named entity may actually refer to two totally different events.

Currently, biologists use their domain knowledge to infer information that text mining cannot predict accurately. Oda *et al.*[72] categorized six inference characters, namely, the state of an entity before or after reaction, the function of an entity before or after a reaction, the influence of state or functional changes of an entity, related reactions, reverse reactions, and characteristics of reactions. If annotated corpora incorporated these features, TMSs would be able to infer information with little human help[72].

Experimental data even confuse biologists sometimes. Open databases of pathway references, such as BioCarta②, STKE③, and KEGG[73], enable biologists to predict the next steps of protein pathways. However, subtle factors cause the results of many experiments to deviate from what is considered consistent for proven pathways. Inconsistencies do not necessarily mean that the proven pathways are wrong, but they may indicate mechanisms or parts of pathways that were previously unobserved. Therefore, pathway prediction should be independent of experiments. In the future, TMSs may be able solve many of the arguments or discrepancies that occur in research today because of their ability to map large amounts of data quickly.

4 Text-Mining Resources

Text-mining resources, such as domain-specific thesauri, lexicons, terminology standards, ontologies, and additional evaluations by task-based challenges are very important. We summarize them in the following subsections. It is our hope that more resources will be used to accelerate progress in the field.

4.1 Evaluating Text Mining via Task-Based Challenges

Evaluation via task-based challenges is essential to the biology community[74]. To date, several biological tasks, including document retrieval, NER, and relation extraction, have been evaluated. We list the major challenges below:

- The KDD Cup 2002 task 1④[75] asked participants to identify papers to be curated for Drosophila gene expression.

- The TREC Genomics Track⑤[76], one of the largest and longest-running challenge evaluations in biomedicine (from 2003 to 2007), evaluates systems for information retrieval.

- The Genic Interaction Extraction⑥ (GIE) challenge[77], a part of the Learning Language in Logic workshop, evaluates the ability of participating TMSs to identify protein/gene interactions from biological abstracts.

- BioCreAtIvE⑦[2] is a community-wide effort that promotes the development and evaluation of text-mining and IE systems applied in the biological domain. The most recent challenge (BioCreAtIvE II.5) in March 2009, which also involved the publisher Elsevier/FEBS Letters and the MINT database, evaluated real-time

② http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways

③ http://stke.sciencemag.org/

④ http://www.biostat.wisc.edu/ craven/kddcup/

⑤ http://ir.ohsu.edu/genomics/

⑥ http://genome.jouy.inra.fr/texte/LLLchallenge/

⑦ http://www.biocreative.org/

text-mining capabilities on full text articles.

- The BioNLP shared task[8] is concerned with the recognition of bio-molecular named entities[1] and events[78] that appear in biomedical literature. The 2009 task used a dataset based on the GENIA event corpus[79]. In contrast to BioCreAtIvE II.5, which aims to support the curation of PPI databases, the BioNLP task concerns to support the development of more detailed and structured databases, e.g., pathway databases[80], and the Gene Ontology Annotation databases[81].

4.2 Text-Mining Corpora

4.2.1 Named Entity Identification Corpora

- The GENIA corpus[9][82] contains 2000 abstracts taken from the MEDLINE database and annotated with various levels of linguistic and semantic information. Biological named entities were annotated according to the taxonomy defined in GENIA ontology. Currently, there are 47 biological named entity categories.
- GENETAG[10][83] is a corpus of 20 000 sentences taken from MEDLINE abstracts annotated with gene/protein names.
- The dataset of the JNLPBA Bio-NER task[11] is annotated with five types of named entities: protein, DNA, RNA, cell line and cell type.
- The training and test sets of BioCreAtIvE I/II gene mention and normalization tasks[12] provide an evaluation standard for the two problems.
- The Yapex Corpus[13] is annotated with protein names mentioned in MEDLINE abstracts related to molecular interaction and published between 1996 and 2001.
- A disease corpus[14] provided by Jimeno *et al.*[34] could serve as a benchmark for other disease NER systems.

4.2.2 Relation Extraction Corpora

- The GENIA event corpus[79] is based on the GENIA corpus and is annotated with events mentioned in biomedical abstracts.
- Binarized BioInfer[15][84] is a corpus annotated with the binary relations between proteins in abstracts.
- AIMed[16][54] is a corpus constructed by using the query word "human" to obtain abstracts from MEDLINE. In total, 1955 sentences were extracted and annotated with gene/protein names and PPIs.
- EDGAR[17][57] contains annotation for the interaction of drugs, genes, and cells.
- The FetchProt[18] corpus is comprised of 190 full text articles of which 140 describe experimental evidence for tyrosine kinase activity in at least one protein. Its annotation includes specific experiments and results, the proteins involved in the experiments and related information.
- The BioText project[19] provides two corpora for relation extraction: 1) PPI data[55] annotates the interaction types between proteins in full texts; and 2) a corpus containing abstracts randomly selected from MEDLINE 2001 for evaluation of mining disease-treatment relations[85].
- The IEPA corpus[20][86] contains 303 PubMed abstracts with annotations for PPIs for each sentence.
- The Craven group's IE datasets[21][56] were compiled from MEDLINE abstracts. There are three datasets, which are labeled, respectively, with instances of the following binary relations: 1) subcellular-localization gathered from the Yeast Proteome Database (YPD); 2) disease-association gathered from the Online Mendelian Inheritance in Man database (OMIM); and 3) PPIs from the MIPS Comprehensive Yeast Genome Database (CYGD).
- The BioCreAtIvE-PPI dataset and DIPPPI corpus[22] were derived from the dataset of BioCreAtIvE I task 1A and the Database of Interaction Proteins (DIP) respectively. The BioCreAtIvE-PPI corpus contains 1000 sentences annotated with PPI information. The PPIs annotated in the DIPPPI corpus are restricted to proteins from yeast. The goal is to find evidence of relations in the text of a paper. Whenever possible, full texts are included in the corpus as

[8] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/
[9] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
[10] ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/GENETAG.tar.gz
[11] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html
[12] http://sourceforge.net/projects/biocreative/files/
[13] http://www.sics.se/humle/projects/prothalt/#data
[14] ftp://ftp.ebi.ac.uk/pub/software/textmining/corpora/diseases
[15] http://mars.cs.utu.fi/BioInfer/
[16] ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/
[17] ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/EDGAR_GS.txt
[18] http://fetchprot.sics.se/#corpus
[19] http://biotext.berkeley.edu/data.html
[20] http://class.ee.iastate.edu/berleant/s/IEPA.htm
[21] http://www.biostat.wisc.edu/~craven/ie/
[22] http://www2.informatik.hu-berlin.de/~hakenber/corpora/

well as abstracts.

- The training and test dataset for the GIE challenge[23][87] contain annotations for gene interactions. Each dataset is decomposed into two subsets. The first subset does not include co/cross-references or ellipsis, but the second subset contains both features.

4.2.3 *Part-of-Speech, Syntactic and Semantic Annotations*

- PASBio[24][88] and BioProp[25][89] contain predicate-argument structures (PAS) for event extraction in molecular biology.
- The PennBioIE[90] CYP corpus[26] contains 1100 PubMed abstracts on the inhibition of cytochrome P450 enzymes. It is annotated with paragraph, sentence boundary, and part-of-speech (POS) information. In addition, 324 of the abstracts are syntactically annotated. Another PennBioIE corpus, the PennBioIE Oncology corpus[27] contains similar annotations but in addition to its abstract is related to cancer concentrating on molecular genetics.
- The GENIA corpus contains annotations for parts-of-speech (POS)[91] and a treebank[92].
- The Brown-GENIA Treebank[28][93] contains the syntactic structures of 21 abstracts (215 sentences) taken from the GENIA corpus. There is no overlap with the GENIA treebank (beta version, 500 abstracts).
- MedPost[94] is a corpus[29] containing 5700 sentences selected randomly from various thematic subsets of MEDLINE and annotated with POS information.
- The PDG Bio-splitter corpus[30]contains a small collection of text datasets compiled from PubMed abstracts to develop sentence splitting tools.
- The BioText project provides a corpus annotated with the definitions of abbreviations[36] taken from 1000 randomly selected abstracts by querying MEDLINE with the term "yeast".

4.2.4 *Full Text Corpora*

- BioMed Central's open access full-text corpus[31] has released 55 003 full text articles to date, including structured XML version, covered by open access license agreements.
- The PPI corpus of BioCreAtIvE II and II.5[95].
- The FlySlip[32] corpus[96] is the first corpus of biomedical full-text articles to be annotated with anaphora information.
- The molecular interaction maps corpus[33][97] contains passages from full-text articles that describe interactions summarized in a molecular interaction map[98].

5 Conclusions

We have considered important research issues related to the exploitation of text mining in the biomedical field, and drawn the following conclusions.

1) The availability of full texts is clearly very important because abstracts usually lack sufficient relevant information. Techniques for mining information from full biomedical texts need to be improved substantially.

2) Text mining has the potential to be used in different applications and to fuse knowledge in the literature and biological databases. However, to realize text mining's full potential, new methods are needed, such as methods for acronym and co-reference resolution, and the integration of various data sources. If highly complex texts and bio-inference sentences can be processed efficiently and accurately, information fusion would enable biologists to exploit knowledge more effectively.

Although text-mining technologies are now quite mature, there are still some important unresolved problems in the field. Fortunately, biomedical text mining is an extremely active research area, and the outlook for continued progress is encouraging. We can foresee that the texts of articles will be systematically mined by computer programs, allowing the interrelation of journal texts and the vast repository of knowledge to be stored semi-automatically in databases. It is expected that text mining tools will be used by every biologist in the future.

References

[1] Kim J D *et al.* Introduction to the bio-entity recognition task at JNLPBA. In *Proc. the International Workshop on Natural Language Processing in Biomedicine and Its Applications* (*JNLPBA2004*), Geneva, Switzerland, Aug. 28-29, 2004, pp.70-75.

[2] Hirschman L *et al.* Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC*

[23] http://genome.jouy.inra.fr/texte/LLLchallenge/#task1

[24] http://research.nii.ac.jp/~collier/projects/PASBio/

[25] http://bws.iis.sinica.edu.tw/BioProp/

[26] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T20

[27] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T21

[28] http://bllip.cs.brown.edu/resources.shtml#corpora

[29] ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz

[30] http://www.pdg.cnb.uam.es/martink/LINKS/biosplitter_corpus.htm

[31] http://www.biomedcentral.com/info/about/datamining/

[32] http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip

[33] http://www.it.usyd.edu.au/~tara/mim_corpus/

Bioinformatics, 2005, 6(Suppl.1): S1.

[3] Krallinger M *et al.* Evaluation of text-mining systems for biology: Overview of the Second BioCreative community challenge. *Genome Biology*, 2008, 9(Suppl. 2): S1.

[4] Hearst M A. Untangling text data mining. In *Proc. the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, College Park, USA, June 20-26, 1999, pp.3-10.

[5] Hahn U *et al.* Text mining: Powering the database revolution. *Nature*, 2007, 448(7150): 130.

[6] Hearst M. What is text mining. 2003, http://people.ischool.berkeley.edu/∼hearst/text-mining.html.

[7] Dai H J *et al.* BIOSMILE web search: A web application for annotating biomedical entities and relations. *Nucl. Acids Res.*, 2008, 36(Web Sever Issue): W390-W398.

[8] Rebholz-Schuhmann D *et al.* Text processing through Web services: Calling Whatizit. *Bioinformatics*, 2008, 24(2): 296-298.

[9] Fernández J M *et al.* iHOP web services. *Nucl. Acids Res.*, 2007, 35(Web Server Issue): W21-W26.

[10] Elsevier Article 2.0 Contest. http://article20.elsevier.com/contest/home.html, Accessed July, 2009.

[11] The Elsevier Grand Challenge. http://www.elseviergrandchallenge.com/, Accessed November, 2009.

[12] BioCreAtIvE II.5. http://www.biocreative.org/events/biocreative-ii5/biocreative-ii5/, Accessed December, 2009.

[13] Ananiadou S, Chruszcz J *et al.* The national ventre for text mining: Aims and objectives. In *Proc. UKKDD2007*, Kent, UK, April 25, 2007, pp.6-12.

[14] RSC Project Prospect. http://www.projectprospect.org/.

[15] Seringhaus M, Gerstein M. Manually structured digital abstracts: A scaffold for automatic text mining. *FEBS Letters*, 2008, 582(8): 1170.

[16] Morgan A *et al.* Overview of BioCreative II gene normalization. *Genome Biology*, 2008, 9(Suppl. 2): S3.

[17] Gonzalez G *et al.* Mining gene-disease relationships from biomedical literature: Weighting protein-protein interactions and connectivity measures. In *Proc. the Pacific Symposium on Biocomputing*, 2007, 12: 28-29.

[18] Tsai R T H, Lai P *et al.* HypertenGene: Extracting key hypertension genes from biomedical literature with position and automatically-generated template features. *BMC Bioinformatics*, 2009, 10(Suppl. 5): S9.

[19] Cohen A M, Hersh W R. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 2005, 6(1): 57-71.

[20] Smith L *et al.* Overview of BioCreative II gene mention recognition. *Genome Biology*, 2008, 9(Suppl.2): S2.

[21] Krallinger M *et al.* Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 2008, 9(Suppl.2): S4.

[22] Chinchor N. MUC-7 named entity task definition (Version 3.5). In *Proc. the 7th Message Understanding Conference*, 1997.

[23] Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 2005, 6(4): 357-369.

[24] Erhardt R A A *et al.* Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, 2006, 11(7/8): 315-325.

[25] Liu H *et al.* A study of abbreviations in MEDLINE abstracts. In *Proc. AMIA Annual Symposium*, San Antonio, USA, Nov. 9-13, 2002, pp.464-468.

[26] Tanabe L, Wilbur W J. Tagging gene and protein names in full text articles. In *Proc. the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain* — Volume 3, Philadelphia, USA, July 11, 2002, pp.9-13.

[27] Tanabe L, Wilbur W J. Tagging gene and protein names in biomedical text. *Bioinformatics*, 2002, 18(8): 1124-1132.

[28] Zhao S. Named entity recognition in biomedical texts using an HMM model. In *Proc. the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, Geneva, Switzerland, Aug. 28-29, 2004, pp.84-87.

[29] Kazama J i *et al.* Tuning support vector machines for biomedical named entity recognition. In *Proc. the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain* — Volume 3, Philadelphia, USA, July 11, 2002, pp.1-8.

[30] Finkel J *et al.* Exploiting context for biomedical entity recognition: From syntax to the web. In *Proc. the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, Geneva, Switzerland, Aug. 28-29, 2004, pp.88-91.

[31] Tsai R T H *et al.* NERBio: Using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics*, 2006, 7(Suppl. 5): S11.

[32] Si L *et al.* Boosting performance of bio-entity recognition by combining results from multiple systems. In *Proc. the 5th International Workshop on Bioinformatics*, Chicago, USA, Aug. 21, 2005, pp.76-83.

[33] Altman R *et al.* Text mining for biology — The way forward: Opinions from leading scientists. *Genome Biology*, 2008, 9(Suppl. 2): S7.

[34] Jimeno A *et al.* Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 2008, 9(Suppl. 3): S3.

[35] Yu H *et al.* Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*, 2002, 9(3): 262-272.

[36] Schwartz A S, Hearst M A. A simple algorithm for identifying abbreviation definitions in biomedical text. *Proc. Pac. Symp. Biocomput.*, 2003, 8: 451-462.

[37] Podowski R *et al.* Suregene, a scalable system for automated term disambiguation of gene and protein names. *Journal of Bioinformatics and Computational Biology*, 2005, 3(3): 743-770.

[38] Hirschman L *et al.* Overview of BioCreAtIvE task 1B: Normalized gene lists. *BMC Bioinformatics*, 2005, 6(Suppl. 1): S11.

[39] Cohen W, Minkov E. A graph-search framework for associating gene identifiers with documents. *BMC Bioinformatics*, 2006, 7: 440.

[40] Leitner F. Comparative community assessments for applied biomedical text mining: BioCreative II challenge and metaservices. In *Intelligent Systems for Molecular Biology* (*ISMB*) *and European Conference on Computational Biology* (*ECCB*), *Highlights Track*, Stockholm, Sweden, June 27-July 2, 2009.

[41] Fundel K, Guttler D *et al.* A simple approach for protein name identification: Prospects and limits. *BMC Bioinformatics*, 2005, 6(Suppl. 1): S15.

[42] Hakenberg J *et al.* Me and my friends: Gene mention normalization with background knowledge. In *Proc. the Second BioCreAtIvE Challenge Evaluation Workshop*, Madrid, Spain, April 23-25, 2007, p.23-25.

[43] Seki K, Javed M. Discovering implicit associations between genes and hereditary diseases. In *Proc. Pac. Symp. Biocomput.*, 2007, 12: 316-327.

[44] Cooper J W, Kershenbaum A. Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics*, 2005, 6: 143.

[45] Shah P K *et al.* Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 2003, 4: 20.

[46] Shatkay H *et al.* Integrating image data into biomedical text categorization. *Bioinformatics*, July 15, 2006, 22(14): e446-e453.

[47] Kou Z *et al.* A stacked graphical model for associating information from text and images in figures. In *Proc. Pac. Symp. Biocomput.*, 2007, 12: 257-268.

[48] Saric J *et al.* Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, March 15, 2006, 22(6): 645-650.

[49] Ono T *et al.* Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, Feb. 2001, 17(2): 155-161.

[50] Kim S *et al.* Kernel approaches for genic interaction extraction. *Bioinformatics*, 2008, 24(1): 118-126.

[51] Bunescu R, Mooney R. Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems*, 2006, 18: 171-178.

[52] Barnickel T *et al.* Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS One*, 2009, 4(7): e6393.

[53] Ramani A *et al.* Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 2005, 6(5): R40.

[54] Bunescu R *et al.* Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 2005, 33(2): 139-155.

[55] Rosario B, Hearst M A. Multi-way relation classification: Application to protein-protein interactions. In *Proc. the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, Oct. 6-8, 2005, pp.732-739.

[56] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In *Proc. the 7th International Conference on Intelligent Systems for Molecular Biology*, Heidelberg, Germany, Aug. 6-10, 1999, pp.77-86.

[57] Rindflesch T C *et al.* EDGAR: Extraction of drugs, genes and relations from the biomedical literature. In *Proc. Pac. Symp. Biocomput.*, 2000, 5: 514-525.

[58] Chun H W *et al.* Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. In *Proc. the Pacific Symposium on Biocomputing,* 2006, 11: 4-15.

[59] Tsai R T H *et al.* HypertenGene: Extracting key hypertension genes from biomedical literature with position and automatically-generated template features. *To appear in BMC Bioinformatics*, 2009.

[60] Miyao Y, Sagae K *et al.* Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, 2008, 25(3): 394-400.

[61] Wong L. PIES, a protein interaction extraction system. In *Proc. Pacific Symposium on Biocomputing*, 2001, 6: 520-531.

[62] Castaño J *et al.* Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution for NLP*, Alicante, Spain, June 3-4, 2002.

[63] Pustejovsky J *et al.* Medstract: Creating large-scale information servers for biomedical libraries. In *Proc. the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, USA, July 11, 2002, pp.85-92.

[64] Nguyen N *et al.* Challenges in pronoun resolution system for biomedical text. In *Proc. the Sixth International Language Resources and Evaluation* (*LREC2008*), Marrakech, Morocco, May 28-30, 2008.

[65] Tsai R T H *et al.* PubMed-EX: A web browser extension to enhance PubMed search with text mining features. *Bioinformatics,* 2009, [Epub ahead of print].

[66] Zhang Z *et al.* Bringing Web 2.0 to bioinformatics. *Brief Bioinform.*, 2009, 10(1): 1-10.

[67] Cheung K *et al.* Semantic Web Approach to Database Integration in the Life Sciences. Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Springer, 2007, pp.11-30.

[68] Dowell R *et al.* The distributed annotation system. *BMC Bioinformatics*, 2001, 2: 7.

[69] O'Reilly T. What is Web 2.0: Design patterns and business models for the next generation of software. 2005, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[70] Mons B *et al.* Calling on a million minds for community annotation in WikiProteins. *Genome Biology*, 2008, 9(5): R89.

[71] Baral C *et al.* CBioC: Beyond a prototype for collaborative annotation of molecular interactions from the literature. In *Proc. Computational Systems Bioinformatics Conference*, 2007, 6: 381-384.

[72] Oda K *et al.* New challenges for text mining: Mapping between text and manually curated pathways. *BMC Bioinformatics*, 2008, 9(Suppl. 3): S5.

[73] Kanehisa M *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 2008, 36(Database Issue): D480-D484.

[74] Hirschman L, Blaschke C. Evaluation of Text Mining in Biology. Text Mining for Biology and Biomedicine, Artech House, 2005, pp.213-245.

[75] Yeh A *et al.* Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles. *ACM SIGKDD Explorations Newsletter*, 2002, 4(2): 87-89.

[76] Hersh W, Voorhees E. TREC genomics special issue overview. *Information Retrieval*, 2009, 12(1): 1-15.

[77] Hakenberg J, Plake C *et al.* LLL'05 challenge: Genic interaction extraction-identification of language patterns based on alignment and finite state automata. In *Proc. the ICML05 Workshop: Learning Language in Logic* (*LLL05*), 2005, 14: 38-45.

[78] Kim J D *et al.* Overview of BioNLP'09 shared task on event extraction. In *Proc. the BioNLP 2009 Workshop Companion Volume for Shared Task*, Boulder, USA, June 4-5, 2009, pp.1-9.

[79] Kim J D *et al.* Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 2008, 9: 10.

[80] Bader G *et al.* Pathguide: A pathway resource list. *Nucleic Acids Research*, 2006, 34(Database Issue): D504-D506.

[81] Camon E *et al.* The gene ontology annotation (GOA) database: Sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 2004, 32(Database Issue): D262-D266.

[82] Kim J D *et al.* GENIA corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics*, 2003, 19(Suppl. 1): 180-182.

[83] Tanabe L *et al.* GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 2005, 6(Suppl. 1): S3.

[84] Heimonen J *et al.* Complex-to-pairwise mapping of biological relationships using a semantic network representation. In *Proc. the Third International Symposium on Semantic Mining in Biomedicine* (*SMBM2008*), Turku, Finland, Sept. 1-3, 2008, pp.45-52.

[85] Rosario B, Hearst M A. Classifying semantic relations in bioscience texts. In *Proc. the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain, July 21-26, 2004, Article No. 43.

[86] Berleant D *et al.* Corpus properties of protein interaction descriptions in MEDLINE. 2003, http://class.ee.iastate.edu/berleant/home/me/cv/papers/corpuspropertiesstart.

htm.

[87] Nedellec C. Learning language in logic-genic interaction extraction challenge. In *Proc. the ICML05 Workshop: Learning Language in Logic* (*LLL05*), Bonn, Germany, Aug. 7, 2005, pp.31-37.

[88] Wattarujeekrit T *et al.* PASBio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, Oct. 19, 2004, 5: 155.

[89] Chou W C *et al.* A semi-automatic method for annotating a biomedical proposition bank. In *Proc. ACL Workshop on Frontiers in Linguistically Annotated Corpora*, Sydney, Australia, July 22, 2006, pp.5-12.

[90] Seth K *et al.* Integrated annotation for biomedical information extraction. In *Proc. HLT/NAACL-2004*, Boston, USA, May 2-7, 2004, pp.61-68.

[91] Tateisi Y, Tsujii J. Part-of-speech annotation of biology research abstracts. In *Proc. the 4th International Conference on Language Resource and Evaluation* (*LREC2004*), Lisbon, Portugal, May 26-28, 2004, pp.1267-1270.

[92] Tateisi Y *et al.* Syntax annotation for the GENIA corpus. In *Proc. IJCNLP 2005*, *Companion Volume*, Jeju Island, Korea, Oct. 11-13, 2005, pp.222-227.

[93] Lease M, Charniak E. Parsing biomedical literature. In *Proc. the Second International Joint Conference on Natural Language Processing*, Jeju Island, Korea, Oct. 11-13, 2005, pp.58-69.

[94] Smith L *et al.* MedPost: A part-of-speech tagger for BioMedical text. *Bioinformatics*, September 22, 2004, 20(14): 2320-2321.

[95] Krallinger M *et al.* The BioCreative II.5 challenge overview. In *Proc. the BioCreative II.5 Workshop 2009 on Digital Annotations*, Madrid, Spain, Oct. 7-9, 2009, p.19.

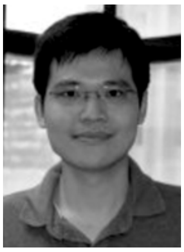
[96] GasperIn C *et al.* Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proc. the Discourse Anaphora and Anaphor Resolution Colloquium*, Lagos (Algarve), Portugal, March 29-30, 2007, pp.19-24.

[97] McIntosh M, Curran J. Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, 2009, 10: 311.

[98] Kohn K W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell*, August 1, 1999, 10(8): 2703-2734.

Hong-Jie Dai received his B.S. degree in computer science and information engineering from Tung Hai University and his M.S. degree in computer science and information engineering from “National Central University in Taiwan” in 2003 and 2005, respectively. Since 2005, he has been an assistant researcher at the “Academia Sinica”. His research interests include: bioinformatics, machine learning, text mining, natural language processing and software engineering.

Yen-Ching Chang received her B.S. degree in biochemical science and technology from “National Taiwan University” and her M.S. degree in biochemistry and molecular biology from “National Taiwan University”, College of Medicine. Since 2008, she has been an research assistant at the “Academia Sinica”. Her research interests include proteomics and text mining.



Richard Tzong-Han Tsai received the B.S. degree in computer science and information engineering from “National Taiwan University”, Taipei, in 1997, the M.S. degree in computer science and information engineering from “National Taiwan University” in 1999, and the Ph.D. degree in computer science and information engineering from “National Taiwan University” in 2006. He was a postdoctoral fellow at “Academia Sinica” from 2006 to 2007. He is now an assistant professor of Department of Computer Science and Engineering, Yuan Ze University, Zhongli, Taiwan, China. His research areas are natural language processing, cross-language information retrieval, biomedical literature mining, and information services on mobile devices.

Wen-Lian Hsu is a distinguished research fellow in the Institute of Information Science, “Academia Sinica”. He received his B.S. degree from the Department of Mathematics, “National Taiwan University” in 1973 and his Ph.D. degree in operations research from Cornell University in 1980, respectively. He then joined Northwestern University, and was promoted to tenured associate professor in 1986. In 1989, he joined the Institute of Information Science as a research fellow. Earlier in his career, Dr. Hsu focused on theoretical graph algorithms and frequently published papers in top-notch journals, such as JACM, SIAM Journal on Computing. After returning to Taiwan, China, he started research on automatic conversion of Pinyin to characters. In 1993, he invented a Chinese natural input method which has since attracted two million users and revolutionized the phonetic input for Chinese in Taiwan, China. Later, he moved into question answering, and bioinformatics. He is currently the director of the International Graduate Program in Bioinformatics in “Academia Sinica”. Dr. Hsu received many awards including the Distinguished Research Fellow Award of the “National Science Council”, K. T. Li breakthrough award, IEEE fellow, “Academia Sinica Investigator” Award, and Teco Technology Award. From 2001 to 2002, he was the President of the Artificial Intelligence Society in Taiwan, China.