# A Survey of State of the Art Biomedical Text Mining Techniques for Semantic Analysis

**Hong-Jie Dai**[1,3]
hongjie@iis.sinica.edu.tw

**Chi-Hsin Huang**[1]
sinyuhgs@iis.sinica.edu.tw

**Jaimie Yi-Wen Lin**[1]
jaimie@iis.sinica.edu.tw

**Pei-Hsuan Chou**[1]
onlytaco@gmail.com

**Richard Tzong-Han Tsai**[2]
thtsai@saturn.yzu.edu.tw

**Wen-Lian Hsu**[1,3*]**, Fellow, IEEE**
hsu@iis.sinica.edu.tw

[1]*Institute of Information Science, Acdemia Sinica, Taipei, Taiwan, R.O.C.*
[2]*Dept. of Computer Science & Engineering, Yuan Ze Univ., Taoyuan, Taiwan, R.O.C.*
[3]*Dept. of Computer Science, National Tsing-Hua Univ., Hsinchu, Taiwan, R.O.C.*

## Abstract

*In recent years, a range of text-mining applications have been developed to improve access to knowledge for biologists and database curators. This paper surveys text-mining works published from 2006 to 2008, with the emphasis on named entity recognition, biological relation extraction and currently available online biological text mining services.*

## 1. Introduction

In a 2004 survey, Cohen [7] et al. observed that the phenomenal growth in biomedical literature poses a major problem for biologists. At present, there are approximately seventeen million articles in the MEDLINE/PubMed database. Clearly, applications that could automatically extract useful information from such massive information sources would greatly facilitate biological research.

The past few years have seen a great deal of research activity in the field of biomedical text mining., The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) [18] task, first held in 2003 and again in 2006, has provided a standard training/evaluation dataset and well defined evaluation metrics for biomedical text mining and information extraction. The task has facilitated cooperation and collaboration between research teams from institutions worldwide and provided a forum for biomedical text mining research. For example, in 2004, The National Centre for Text Mining (NaCTeM) [2] was established by the University of Manchester, the University of Liverpool and the University of the Salford with the objective of offering high quality biological and biomedical text mining services. Many other projects have sprung up in the wake of BioCreAtIvE [18]. A survey conducted by Alex et al. [1] of the latest biomedical natural language processing (NLP) technology showed that a maximum reduction of one-third in curation time can be expected, showing that biomedical text mining is a promising field.

In this paper, we present a survey of recent biomedical text mining works published between the end of 2006 and the beginning of 2008. The survey covers 13 openly available named entity recognition (NER) systems, four semantic role labeling (SRL) corpora, two event corpora, and 12 text mining-based web services.

## 2. Biological Named Entity Recognition

The first step in biological text-mining is the identification of biological entities, referred to as the NER task. NER is important because it is a fundamental step for tasks, such as information extraction, summarization, and question answering. In the biological realm, the types of named entities (NEs)

[*] Corresponding author

are wider in scope than the generic entity types PERSON, ORGANIZATION and LOCATION defined in [5]. In 2004, the JNLPBA [20] open challenge task for bio-NER simplified the 36 entity classes in the GENIA corpus [21] and used only five classes, namely protein, DNA, RNA, cell line, and cell type, to evaluate the performance of the participating systems. Unlike the earliest rule-based NER system [14], the following four types of classification models were applied by the participating teams: Support Vector Machines (SVMs) [16, 34], Hidden Markov Models (HMMs) [50], Maximum Entropy Markov Models (MEMMs) [13] and Conditional Random Fields (CRFs) [38]. The most frequently applied models were SVMs. The evaluation results showed that SVMs worked better in combination with other models, while the other three models yielded a reasonable performance in isolation [20]. However, the CRFs system proposed by Settles [38] achieved a comparable performance to that of the top ranked systems [16] with a simple feature set, which suggests that integration of more useful features may further improve the NER performance.

In 2006, the second BioCreAtIvE workshop organized a gene mention tagging task [42], which involved 21 teams. In contrast to JNLPBA 2004, half of the teams used CRFs as their machine-learning models, and almost all participating teams used machine-learning-based approaches. This indicates that, since annotated corpora became available, machine-learning approaches have become the mainstream for NER tasks [51]. Specifically, the most popular model is CRFs [9, 27, 30, 44].

One contribution of the second BioCreAtIvE workshop was the launch of the BioCreative MetaServer (BCMS) online web service, which integrates about twenty annotation servers in different countries to provide NER annotation services. Users or computer programs can simply access the service via BCMS's uniform application programming interfaces without considering the fact that the annotation results derive from different annotation servers using a variety approaches. The scalability (number of participants) of BCMS is its main advantage.

Some participating teams made their gene mention tagging tools openly available [9, 19]. NERBio [9] and AIIAGMT [19], which are both based on CRFs, are easy-to-use online tools for detecting gene and gene product names in free text. NERBio applies the numerical normalization technique [44] to substantially reduce the number of features required for machine-learning training, and also to improve the accuracy of feature weight estimation. Numerical normalization is useful because entity names often occur in a series, such as the gene names IL2, IL3, and so on. The numeric normalized value for them is IL0; hence, the unseen surface forms, such as IL4, in the training data have the same representation as forms that are seen. AIIAGMT combines the tagging results with forward and backward parsing to improve its performance [26].

In addition to the above online NER services, three downloadable tools, Penn BioTagger [15], GENIA Tagger [48] and BANNER [28], have been released. Penn BioTagger was trained by using the k-best MIRA learning algorithm [29] with lexicons and automatically derived word clusters. It achieved a final F-measure of 86.28% (ranked 5 in the second BioCreAtIvE workshop). Originally, the GENIA Tagger only output base forms, part-of-speech tags and chunk tags, but the latest version, GENIA Tagger 3.0, also supports NE tags. BANNER is an open-source bio-NER tool that uses CRFs with carefully selected feature sets and the numerical normalization technique [44]. The evaluation results [28] show that BANNER yields a significantly better performance than existing open source systems, including ABNER [39] and LingPipe [3]. Because BANNER is open-source, it can be re-trained with new NER corpora; hence, researchers who require a baseline system can use it as a benchmark for evaluating new methods they propose for NER tasks.

In contrast, the BioCaster text mining project[2], which is dedicated to the detection and tracking of disease outbreaks from Internet news articles, provides a totally different perspective of NER problems. The above NER tasks defined several annotation schemas, such as DNA and RNA, for biomedical text; however, until recently, little work had been done on developing a schema specifically for public health related texts. In 2007, Doan et al. [10] of the BioCaster project developed an annotation schema to fill this research gap. They identified several important concepts that reflect information about infectious diseases, and created guidelines for annotating them as target NE classes. In total, 18 concepts are specified as NE classes, namely PERSON, LOCATION, ORGANIZATION, TIME, DISEASE, CONDITION, OUTBREAK, VIRUS, ANATOMY, PRODUCT, NONHUMAN, DNA, RNA, PROTEIN, CONTROL, BACTERIA, CHEMICAL, and SYMPTOM. After defining the 18 categories, Doan et al.

[2] http:*//biocaster.nii.ac.jp*

**Table 1. Openly available NER tools or services**

| Name | Description | URL |
|---|---|---|
| AbbreviationServer [5] | Biomedical abbreviation server | http://bionlp.stanford.edu/abbreviation/ |
| AbGene [50] | Protein name tagger | ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe |
| ABNER [46] | Protein/Gene/DNA/RNA/cell tagger | http://pages.cs.wisc.edu/~bsettles/abner/ |
| AIIAGMT [22] | Gene and protein name tagger | http://140.109.23.113/AIIAGMT/index.html |
| AliasServer [23] | Protein alias handler | http://cbi.labri.fr/outils/alias/index.php |
| BANNER [33] | Gene and protein name tagger | http://banner.sourceforge.net/ |
| BioCaster [13] | Health protection roles tagger | http://biocaster.nii.ac.jp/ |
| BCMS | Gene and protein name tagger | http://bcms.bioinfo.cnio.es |
| GAPSCORE [6] | Protein name tagger | http://bionlp.stanford.edu/gapscore |
| GENIA Tagger [56] | Protein/Gene/DNA/RNA/cell tagger | http://text0.mib.man.ac.uk/software/geniatagger/ |
| NERBio [12] | Gene and protein name tagger | http://asqa.iis.sinica.edu.tw/biocreative2/ |
| NLPort Tagger [36] | Protein name tagger | http://cubic.bioc.columbia.edu/services/NLProt/ |
| Penn BioTagger [18] | Gene and protein name tagger | http://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger.html |

trained the BioCaster tagger with the Näive Bayes classifier [31].

We reviewed nine NER tools described in [25], and summarize all currently available NER tools in Table 1.

## 3. Biological Relation Corpora

In this section, we shift our focus from the fundamental NER task to the task of extracting verbal information that represents the relations between NEs.

The simplest way to detect the relations between NEs is to collect texts in which they co-occur. In most cases, co-occurrence statistics provide high recall but poor precision, but they can often be used as a baseline system against which other methods can be compared [16]. Advanced approaches that determine the roles played by NEs can be roughly classified into three categories. (1) Pattern-based methods, which map words, parts-of-speech, or NEs sequences into structural information slots according to predefined patterns and matching rules [32-34]. (2) Natural language processing based methods, which may use full parsing or shallow parsing information to extract subject/object information from predefined frames [35, 36]. Huang et al. [37] proposed using a hybrid method with both shallow parsing and pattern matching. A completely different technique that utilizes a Web search engine was proposed by Mukherjea et al. [32]. (3) The semantic role labeling (SRL) technique, which we discuss in detail below.

In SRL, sentences are represented by one or more predicate-argument structures (PASs), also known as propositions [33]. Each PAS is composed of a predicate (e.g., a verb) and several arguments (e.g., noun phrases and adverbial phrases) that have different semantic roles. The roles include main arguments, such as an agent and a patient, as well as adjunct arguments, such as time, manner, and location. In 2004, the PASBio [49] project released a set of PASs for a small set of biometrically relevant verbs. PASBio is specifically designed for annotating molecular events and defining core arguments that are important for completing the meaning of an event.

Because the PASBio project only focused on the creation of a semantic lexicon and annotation guidelines, some researchers have extended it to create useful biomedical applications. For instance, Kogan et al. [23] extended PASBio to build a domain-specific set of PASs for the medical domain, while Shah et al. [40] used the PASBio's representation scheme to construct semantic patterns for the LSAT (Literature Support for Alternative Transcripts) database system. In 2006, Shah et al. [41] annotated a small PASBio corpora to build a semantic role labelling system. They showed that a prior binary classification step could constrain the number of predicates, and provided greater insight into the semantic roles of sentence constituents for biomedical event extraction. These successful applications show that the PASBio method and its specific representational schemes are adequate for the general problem of representing molecular biology concepts [8].

In 2006, Chou et al. [6] proposed another realizable approach for constructing a biomedical proposition bank on top of GENIA Treebank (GTB) [43]. To construct their biomedical proposition bank, they first employed the rich resources of PropBank [33] in the general English domain to build an SRL system [47]. They then used the SRL system to automatically annotate the semantic roles in GTB and construct the biomedical proposition bank called BioProp [6]. The project involved annotating the arguments of 30 frequent biomedical verbs. In contrast to PASBio, BioProp does not place any biomedical constraints on

**Table 2. Biological Relation corpora**

| Name | Description | URL |
|---|---|---|
| BioInfer [42] | A biological relationships corpus | http://mars.cs.utu.fi/BioInfer/ |
| GENIA event corpus [26] | A biological event annotation corpus | http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Event+Annotation |
| Kogan et al. [27] | A medical domain SRL corpus | http://ycmi.med.yale.edu/krauthammer/rolelabeling.htm |
| LSAT [48] | Literature Support for Alternative Transcripts | http://www.bork.embl.de/LSAT/ |
| PASBio [57] | A set of PAS for semantic roles of biomedical verbs | http://research.nii.ac.jp/~collier/projects/PASBio/ |

its PASs because of the PropBank standard. Furthermore, BioProp provides complete structures for describing argument modifiers, such as location, manner, timing, and condition. The primary goal of BioProp is to port the proposition bank to the biomedical domain for training a biomedical SRL system called BIOmedical SeMantIc roLe labEler (BIOSMILE)[45].

In addition to PASBio and BioProp, another corpus called Bioinfer was released in  2007 [35]. Bioinfer is annotated with syntactic dependencies and NEs as well as their relationships within a complex structure, such as relationships between relationships or the relationships of more than two entities. Ontologies that define the types of entities and relationships annotated in the corpus are also provided. Currently, the corpus contains 1,100 sentences from abstracts of biomedical research articles.

The latest GENIA event corpus [22] was released in January 2008. A new type of annotation, called event annotation, has been added to the corpus. Event annotation belongs to what we call biological annotation. In contrast to linguistic annotation, such as SRL discussed earlier, biological annotation is performed by biologists, not by linguists. It follows a similar principle to that used in the annotation of Bioinfer, i.e., it associates all annotations with actual expressions in text. The difference between the two types of annotation is that the goal of biological annotation is to identify what kinds of biological information appear in which part of the text, while linguistic annotation focuses on the linguistic properties of texts in the domain. NE annotation in the GENIA event corpus is one example of biological annotation. It identifies text spans in which biological entities, such as proteins, DNA, RNA, and cellular locations actually appear. This new annotation was made on half of the GENIA corpus [21], consisting of 1,000 Medline abstracts. The GENIA event corpus contains 9,372 sentences in which 36,114 events have been identified.

We summarize the current openly available corpora in Table 2.

## 4. Biological Web Services

From the large number of publications in the biological text mining area, it is clear that the performance of basic text mining tasks has reached reasonable levels. In the last decade, several advanced biological text-mining services have been developed, and some systems have been applied to real-world curation problems. PreBIND [11], for example, was developed to facilitate the extraction of protein-protein interactions (PPI) and reduces the task duration by 70% [11]. The PRIME [24] database text mining system, on the other hand, extracts interactions between proteins, genes and compounds. A new text mining system, EpiLoc [4], which predicts the subcellular location of proteins was published at the beginning of 2008. It applies subcellular localization prediction to almost any protein, even in the absence of published data about it.

For article retrieval, biologists are now able to search through a massive volume of online articles. For example, using NCBI PubMed Entrez [37], a user can retrieve articles from a database of over 4,600 biomedical journals published from 1966 to the present; the database is updated daily. BioText [17] provides a new way to access scientific literature by enabling biologists to search and browse the figures and captions in biological articles. However, users of these basic search engines may need to scan or read retrieved articles in more detail to obtain specific information of interest. Needless to say, services that can identify and mark key relations, entities and terms can save biologists a great deal of time.

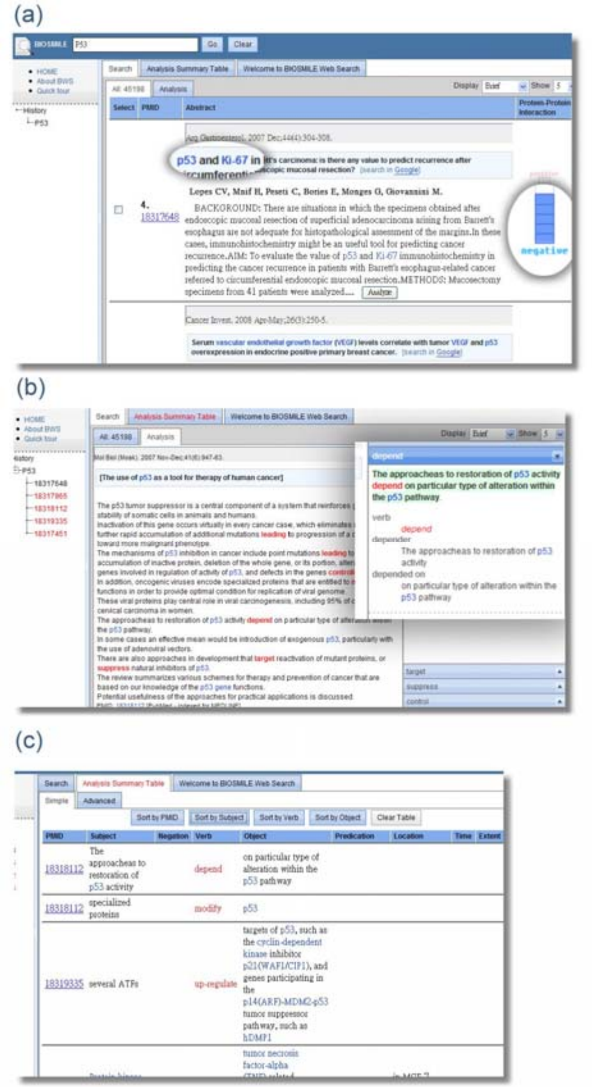
Several advanced search services have already been developed. For example, BESearch [46] provides biologists with a form-based query interface to obtain the information they need. Meanwhile, the iHOP service [12] retrieves sentences containing specified genes, labels the biomedical entities in the genes, and provides graphs of the co-occurrences among all entities. iHOP allows researchers to (1) filter and rank retrieved sentences that match the given gene or protein names according to their significance, impact factor, date of publication and syntax; and (2) explore a network of gene and protein interactions by directly navigating the pool of published scientific literature. MEDIE, developed by the Tsujii Laboratory, can identify subject-verb-object (syntactic) relations and biomedical entities in sentences.

Another novel text mining service called BIOSMILE Web Search (BWS) was released in February 2008. BWS has similar features to iHOP and MEDIE. It can annotate entities as well as a wider range of relation types (Figure 1). For example, the sentence "KaiC enhanced KaiA-KaiB interaction in vitro and in yeast cells," describes an enhancement relation. BWS can identify the elements in this relation, such as the action "enhanced", the enhancer "KaiC", the enhanced "KaiA-KaiB interaction", and the location "in vitro and in yeast cells". Relations are classified by their main verbs and put in different tabs. This makes it easy for researchers to browse through all the relations in an article verb by verb, helps them locate passages of interest easily, and significantly speeds up overall comprehension (see Figure 1b). BWS also provides a search result summary in table format, showing all the relations found in multiple articles during one session (see Figure 1c). This is a convenient function for summarizing several related papers. Furthermore, for researchers interested in PPI, BWS classifies articles as PPI-relevant or –irrelevant [36].

We summarize the current biological web services in Table 3.

## 5. Conclusion

As the goals of biomedical information extraction applications have become more ambitious, the range of bio-NLP application types has become correspondingly broader. In this paper, we have summarized state of the art bio-NLP applications ranging from fundamental NER to more complex relation extraction and online integrated text mining services. Needless to say, there are still significant unsolved problems in the field. However, biomedical text mining is an extremely active research area, and the outlook for continued progress is positive.



Arq Gastroenterol. 2007 Dec;44(4):304-308.

p53 and Ki-67 in Barrett's carcinoma: is there any value to predict recurrence after circumferential endoscopic mucosal resection? [search in Google]

Lopes CV, Maif H, Pesati C, Bories E, Monges G, Giovannini M.

4. 18317648 BACKGROUND: There are situations in which the specimens obtained after endoscopic mucosal resection of superficial adenocarcinoma arising from Barrett's esophagus are not adequate for histopathological assessment of the margins.In these cases, immunohistochemistry might be an useful tool for predicting cancer recurrence.AIM: To evaluate the value of p53 and Ki-67 immunohistochemistry in predicting the cancer recurrence in patients with Barrett's esophagus-related cancer referred to circumferential endoscopic mucosal resection.METHODS: Mucosectomy specimens from 41 patients were analyzed....

Cancer Invest. 2008 Apr-May;26(3):250-5.

Serum vascular endothelial growth factor (VEGF) levels correlate with tumor VEGF and p53 overexpression in endocrine positive primary breast cancer. [search in Google]

Mol Biol (Mosk). 2007 Nov-Dec;41(6):947-63.

[The use of p53 as a tool for therapy of human cancer]

The p53 tumor suppressor is a central component of a system that reinforces stability of somatic cells in animals and humans.
Inactivation of this gene occurs virtually in every cancer case, which eliminates further rapid accumulation of additional mutations leading to progression of a toward more malignant phenotype.
The mechanisms of p53 inhibition in cancer include point mutations leading to accumulation of inactive protein, deletion of the whole gene, or its portion, genes involved in regulation of activity of p53, and defects in the genes control
In addition, oncogenic viruses encode specialized proteins that are entitled to functions in order to provide optimal condition for replication of viral genome.
These viral proteins play central role in viral carcinogenesis, including 95% of cervical carcinoma in women.
The approacheas to restoration of p53 activity depend on particular type of alteration within the p53 pathway.
In some cases an effective mean would be introduction of exogenous p53, particularly with the use of adenoviral vectors.
There are also approaches in development that target reactivation of mutant proteins, or suppress natural inhibitors of p53.
The review summarizes various schemes for therapy and prevention of cancer that are based on our knowledge of the p53 gene functions.
Potential usefulness of the approaches for practical applications is discussed.

The approacheas to restoration of p53 activity depend on particular type of alteration within the p53 pathway.

verb
depend
depender
The approacheas to restoration of p53 activity
depended on
on particular type of alteration within the p53 pathway

| PMID | Subject | Negation | Verb | Object | Predication | Location | Time | Extent |
|---|---|---|---|---|---|---|---|---|
| 18318112 | The approacheas to restoration of p53 activity | | depend | on particular type of alteration within the p53 pathway | | | | |
| 18318112 | specialized proteins | | modify | p53 | | | | |
| 18319335 | several ATFs | | up-regulate | targets of p53, such as the cyclin-dependent kinase inhibitor p21(WAF1/CIP1), and genes participating in the p14(ARF)-MDM2-p53 tumor suppressor pathway, such as hDMP1 | | | | |
| | Protein kinase | | | tumor necrosis factor-alpha (TNF)-related | | in MCF-7 | | |

**Figure 1.** The features of the BWS search interface. (a) Users can enter either a PMID or keywords. For each abstract, BWS annotates gene or protein names in light blue, and a graduated bar meter indicates the abstract's relevance to PPI. (b) Analysis results are shown in the tab pane with biomedical verbs marked in red. The semantic roles related to a verb are listed on the right-hand side. (c) An analysis summary table that contains all relations in abstracts.

**Table 3. Biological Web Services**

| Name | Description | URL |
|---|---|---|
| BIOSMILE Web Search | Biomedical relation extraction service | http://bioservices.cse.yzu.edu.tw/BWS/ |
| BioText [20] | Scientific literature figures and captions search engine | http://biosearch.berkeley.edu/ |
| Chilibot [7] | Relationships search engine | http://www.chilibot.net/ |
| EpiLoc | Subcellular localization prediction system | http://epiloc.cs.queensu.ca |
| iHOP [15] | Information on hyperlinked proteins | http://www.ihop-net.org/ |
| MEDIE | Syntactic relations extraction system | http://www-tsujii.is.s.u-tokyo.ac.jp/medie/ |
| KinasePathway database [28] | Tool for extraction of protein, gene and compound interactions from text | http://kinasedb.ontology.ims.u-tokyo.ac.jp:8081/ |
| PreBIND [14] | Classifier of protein interaction documents | http://bond.unleashedinformatics.com/ |
| PRIME [29] | Tool for extraction of protein, gene and compound interactions from text | http://prime.ontology.ims.u-tokyo.ac.jp:8081/ |
| PubMed Entrez [44] | Biomedical citation retrieval system | http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed |
| Textpresso [39] | *C. elegans* literature information retrieval and extraction tool | http://www.textpresso.org/ |

## 6. References

[1] B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, and X. Wang, "ASSISTED CURATION: DOES TEXT MINING REALLY HELP?," *Pac Symp Biocomput*, vol. 556, p. 67, 2008.

[2] S. Ananiadou, J. Chruszcz, J. Keane, J. McNaught, and P. Watry, "The National Centre for Text Mining: Aims and Objectives," *UKKDD'5*, 2007.

[3] B. Baldwin and B. Carpenter, "LingPipe," in *http://www.alias-i.com/lingpipe/*.

[4] S. Brady and H. Shatkay, "EpiLoc: a (working) text-based system for predicting protein subcellular location," *Pac Symp Biocomput*, vol. 604, p. 15, 2008.

[5] N. Chinchor, "MUC-7 named entity task definition," *Proceedings of the 7th Message Understanding Conference*, 1997.

[6] W. C. Chou, R. T. H. Tsai, Y. S. Su, W. Ku, T. Y. Sung, and W. L. Hsu, "A Semi-Automatic Method for Annotating a Biomedical Proposition Bank," *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*, pp. 5-12, 2006.

[7] K. B. Cohen and L. Hunter, "Natural Language Processing and Systems Biology," *Artificial Intelligence Methods and Tools for Systems Biology*, 2004.

[8] K. B. Cohen and L. Hunter, "A critical review of PASBio's argument structures for biomedical verbs," *BMC Bioinformatics*, vol. 7 Suppl 3, p. S5, 2006.

[9] H.-J. Dai, H.-C. Hung, R. T.-H. Tsai, and W.-L. Hsu, "IASL Systems in the Gene Mention Tagging Task and Protein Interaction Article Sub-task," in *Proceedings of Second BioCreAtIvE Challenge Workshop*, 2007.

[10] S. Doan, A. Kawazoe, and N. Collier, "The Role of Roles in Classifying Annotated Biomedical Text," in *BioNLP 2007*, 2007.

[11] I. Donaldson, J. Martin, B. de Bruijn, and C. Wolting, "PreBIND and Textomy-mining the biomedical literature for proteinprotein," 2003.

[12] J. M. Fernandez, R. Hoffmann, and A. Valencia, "iHOP web services," *Nucl. Acids Res.*, vol. 35, 2007.

[13] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web," *Joint Workshop on Natural Language Processing in Biomedicine and Its Applications at Coling 2004*, 2004.

[14] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: identifying protein names from biological papers," *Pacific Symposium on Biocomputing*, pp. 707-718, 1998.

[15] K. Ganchev, K. Crammer, F. Pereira, G. Mann, K. Bellare, A. McCallum, S. Carroll, Y. Jin, and P. White, "Penn/UMass/CHOP Biocreative II systems," in *Proceedings of Second BioCreAtIvE Challenge Workshop*, 2007.

[16] Z. GuoDong, S. Jian, N. Collier, P. Ruch, and A. Nazarenko, "Exploring Deep Knowledge Resources in Biomedical Name Recognition," *COLING 2004 International Joint workshop*

*on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pp. 99-102, 2004.

[17] M. A. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. A. Wooldridge, and J. Ye, "BioText Search Engine: beyond abstract search," *Bioinformatics*, vol. 23, p. 2196, 2007.

[18] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology," *feedback*, 2005.

[19] H. S. Huang, Y. S. Lin, K. T. Lin, C. J. Kuo, Y. M. Chang, B. H. Yang, C. N. Hsu, and I. F. Chung, "High-Recall Gene Mention Recognition by Unification of Multiple Backward Parsing Models," *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, p. 109?11, 2007.

[20] K. Jin-Dong, O. Tomoko, Y. T. Yoshimasa Tsuruoka, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," *Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04)*, p. 70?5, 2004.

[21] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus--a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, pp. 180-182, 2003.

[22] J.-D. Kim, T. Ohta, and J. i. Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC Bioinformatics*, vol. 9:10, 2008.

[23] Y. Kogan, N. Collier, S. Pakhomov, and M. Krauthammer, "Towards Semantic Role Labeling & IE in the Medical Literature," *AMIA Annual Symposium Proceedings*, vol. 2005, p. 410, 2005.

[24] A. Koike, Y. Niwa, and T. Takagi, "Automatic extraction of gene/protein biological functions from biomedical text," *Bioinformatics*, vol. 21, pp. 1227-1236, 2005.

[25] M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology," *Genome Biology*, 2005.

[26] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," *North American Chapter Of The Association For Computational Linguistics*, pp. 1-8, 2001.

[27] C. J. Kuo, Y. M. Chang, H. S. Huang, K. T. Lin, B. H. Yang, Y. S. Lin, C. N. Hsu, and I. F. Chung, "Rich Feature Set, Unification of Bidirectional Parsing and Dictionary Filtering for High F-Score Gene Mention Tagging," *Sumbitted to Second BioCreAtIvE Challenge Workshop*, 2007.

[28] R. Leaman and G. Gonzalez, "BANNER: AN EXECUTABLE SURVEY OF ADVANCES IN BIOMEDICAL NAMED ENTITY RECOGNITION," *Pac Symp Biocomput*, vol. 652, p. 63, 2008.

[29] R. McDonald, K. Crammer, and F. Pereira, "Online Large-Margin Training of Dependency Parsers," *Ann Arbor*, vol. 100, 2005.

[30] R. McDonald and F. Pereira, "Identifying gene and protein mentions in text using conditional random fields.," *BMC Bioinformatics*, vol. 6, p. (Suppl)(1:S6), 2005.

[31] T. M. Mitchell, *Machine Learning*: McGraw-Hill, 1997.

[32] S. Mukherjea and S. Sahay, "DISCOVERING BIOMEDICAL RELATIONS UTILIZING THE WORLD-WIDE WEB," *Pac Symp Biocomput*, vol. 11, pp. 164-75, 2006.

[33] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational Linguistics*, vol. 31, pp. 71-106, 2005.

[34] K. M. Park, S. H. Kim, D. G. Lee, and H. C. Rim, "Boosting Lexical Knowledge for Biomedical Named Entity Recognition," *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, p. 75?9, 2004.

[35] S. Pyysalo, F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski, "BioInfer: a corpus for information extraction in the biomedical domain," *BMC Bioinformatics*, vol. 8, p. 50, 2007.

[36] T. RT-H, H. H-C, D. H-J, and H. W-L, "Exploiting Likely-Positive and Unlabeled Data to Improve the Identification of Protein-Protein Interaction Articles," *6th InCoB - Sixth International Conference on Bioinformatics*, 2007.

[37] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, "Entrez: molecular biology database and retrieval system," *Methods Enzymol*, vol. 266, pp. 141-62, 1996.

[38] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets," in *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its*

*Applications (JNLPBA-2004)* Geneva, Switzerland, 2004.

[39] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text." vol. 21: Oxford Univ Press, 2005, pp. 3191-3192.

[40] P. K. Shah, L. J. Jensen, S. Boue, and P. Bork, "Extraction of transcript diversity from scientific literature," *PLoS Computational Biology,* 2005.

[41] P. K. Shah and P. Bork, "LSAT: learning about alternative transcripts in MEDLINE," *Bioinformatics*, vol. 22, pp. 857-865, 2006.

[42] L. Smith, L. K. Tanabe, R. J. n. Ando, C.-J. Juo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. B. Jr., L. Hunter, B. Carpenter, R. T.-H. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. T. Perez, M. Neves, P. Nakov, A. Divoli, M. Mana, J. Mata-Vazquez, and W. J. Wilbur., "Overview of BioCreative II Gene Mention Recognition," *Genome Biology,* 2007.

[43] Y. Tateisi, A. Yakushiji, T. Ohta, and J. Tsujii, "Syntax Annotation for the GENIA corpus," *Proc. IJCNLP 2005, Companion volume,* pp. 222–227, 2005.

[44] R. T.-H. Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu, "NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition," *BMC Bioinformatics*, vol. 7 Suppl 5, p. S11, 2006.

[45] R. T.-H. Tsai, W.-C. Chou, Y.-S. Su, Y.-C. Lin, C.-L. Sung, H.-J. Dai, I. T. Yeh, W. Ku, T.-Y. Sung, and W.-L. Hsu, "BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features," *BMC Bioinformatics,* vol. 8, p. 325, 2007.

[46] R. T. H. Tsai, H. J. Dai, H. C. Hung, R. T. K. Lin, W. C. Chou, Y. S. Su, M. Y. Day, and W. L. Hsu, "BESearch: A Supervised Learning Approach to Search for Molecular Event Participants," *Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on,* pp. 412-417, 2007.

[47] T. H. Tsai, C. W. Wu, Y. C. Lin, and W. L. Hsu, "Exploiting Full Parsing Information to Label Semantic Roles Using an Ensemble of ME and SVM via Integer Linear Programming," *Proceedings of CoNLL-2005,* 2005.

[48] Y. Tsuruoka, Y. Tateishi, J. D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a robust part-of-speech tagger for biomedical text," *Lecture notes in computer science*, pp. 382-392, 2005.

[49] T. Wattarujeekrit, P. K. Shah, and N. Collier, "PASBio: predicate-argument structures for event extraction in molecular biology," *BMC Bioinformatics*, vol. 5, p. 155, Oct 19 2004.

[50] S. Zhao, "Named Entity Recognition in Biomedical Texts using an HMM Model," *Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA),* 2004.

[51] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, "Frontiers of biomedical text mining: current progress," *Briefings in Bioinformatics,* 2007.