

SysBio Explorer — a Systems Biology Literature Retrieval and Processing Framework

No Author Given

No Institute Given

Abstract. This paper details the *SysBio Explorer*, a Systems Biology Literature Retrieval and Processing Framework, aiming at the automatic inference of regulatory and metabolic networks based on published information. The *SysBio Explorer* does not focus on any organism or problem in particular and encompasses extensible processing and analysis techniques. It works over full-text documents, applying Natural Language Processing techniques and using biomedical dictionaries and ontologies together with hand-made rules. Besides biological entity recognition and relation extraction, document classification, relevance assessment and reference/authoring networks are within its present scope.

The framework is described in terms of its design requirements and implementation decisions, exposing current achievements and future work and also highlighting present obstacles. Experiments over real-world problems concerning the organisms *E. coli*, *S. cerevisiae* and *H. pylori* are used in its validation.

1 Introduction

Biomedical Text Mining (BTM), i.e., the field that deals with the automatic retrieval and processing of biomedical literature, is perhaps one of today's most promising research fields [9]. The large diversity of data to be collected, the heterogeneity of the data sources and the ever growing rate of publication strongly demand for specialised and automated processes. Researchers spend a lot of time and effort in searching for the available information about their particular area of research. Manual curation implies an additional effort, delaying information availability and thus leading to erroneous, resource and time-consuming decisions.

Currently, BTM is still far from sustaining the full automation of the curation procedures, but the achieved breakthroughs are already worth of notice. Mining techniques have been addressing, among others, the tasks of *Named Entity Recognition* (NER), *Relation Extraction* (RE), document summarisation, document classification, document clustering, and abbreviation and synonym resolution. Yet, BTM has to face a major challenge: biomedical terminology. Biomedical terminology is not standardised, and term ambiguity and variation make it very hard to accurately identify mentions to relevant entities and thus, proceed with further information extraction.

Dictionaries, gazetteers (lists of look-up strings) and hand-made rules do not encompass terminology at its full extent and ontologies can only provide partial coverage of the domain. Nevertheless, current biomedical ontologies present a comprehensive body of knowledge that BTM applications can not ignore. Available ontologies together with linguistic and user-specified data can aid in the semantic interpretation of biomedical publications, enhancing BTM processes and even sustaining further update of the ontologies.

The proposed work tackles *Systems Biology* (SB) literature retrieval and processing, in particular, the automatic modelling of regulatory and metabolic networks, combining available ontologies and state-of-the-art BTM techniques. So far, most TM efforts have focused on abstract compilation and processing, specifically NER and, more recently, RE, namely, the discovery of *Protein-Protein Interactions* (PPIs). However, most often, abstracts do not contain the desired regulatory and metabolic data, forcing BTM to encompass full-text processing and analysis.

In this sense, our SB Literature Retrieval and Processing Framework, the *SysBio Explorer*, has two main conceptual goals: (i) to apply BTM techniques to the search of metabolic and regulatory data and (ii) to provide means of automated curation of real-world, user-specified problems. Its design requirements include full-text processing, the conciliation of multiple ontologies, the specification of relation verbs, the XML annotation with parameter specification and document classification based on the set of annotated entities. Besides NER and RE, document classification, relevance assessment and authoring networks (linking researchers to document and mentioned entities) are within its present scope, although further BTM efforts are also devised.

The paper details experiments over real-world problems concerning the organisms *E.coli*, *S.cerevisiae* and *H.pylori*. Such experiments aim at demonstrating the usability and usefulness of the framework without disregarding its present limitations, both in terms of ontology management, text processing and analysis.

2 Named Entity Recognition

NER is a crucial task in any BTM process which can be accomplished using pattern matching and Machine Learning (ML). Dictionaries, hand-made rules and gazetteers are usually used in pattern matching [5], while hidden Markov models [7], naive Bayes, maximum entropy, conditional random field [14], support vector machines [4], decision trees and combinations of heuristics are common ML approaches.

In spite of the chosen approach, knowledge acquisition is necessary. The manual curation of a representative number of documents for a particular problem is time-consuming, but annotated corpora are biased to a particular domain/problem and cannot provide decision models that perform well in user-specified problems. Gazetteers and simplistic dictionaries are not able to encompass sophisticated, detailed terminology. Sophisticated dictionaries and ontologies demand permanent maintenance, and hand-made rules cannot face term

variance and ambiguity properly. Semantically annotated corpora are built from the results of particular keyword-based queries. For example, the well-known GENIA corpus [6], contains documents related to *Human, Blood Cells*, and *Transcription Factors*, i.e., human blood cell transcription factors. Such biased annotation resources are suitable for technique benchmarking, but cannot be used on general, user-specified problem annotation.

The combination of relevant resources is perhaps the most reasonable approach towards general biomedical NER [8,15,16]. Corpora can aid on ML technique development while encyclopedic information grants real-world appliance. Together, lexicon, ontologies and rules may cope with domain's specificities, ensuring support to user-specified problems and term normalisation (mapping text occurrences to well-defined entities). On the other hand, ML techniques may address further analysis of the annotated documents.

3 The SysBio Explorer

The *SysBio Explorer* targets problems related to SB research areas, in particular, aiming at the discovery and processing of regulatory and metabolic data and the subsequent modelling of the corresponding networks. Besides supporting state-of-the-art BTM, this framework differs from existing works because it is meant for common use by Biology researchers without specific tutoring. It provides the means to take into advantage as much information resources as available, detaching from particular organisms or problems. Furthermore, it tackles full-text document in order to extract as much information as possible, facing associated processing issues and extending current techniques.

The design requirements that have guided the development of the framework are the following:

- the search of bibliographic databases, namely MEDLINE's PubMed, in order to collect potentially relevant documents on a user-defined problem (by keyword match), and the actual retrieval of the full texts whenever open access is granted;
- the conversion of PDF documents into plain text;
- the use (conciliation) of multiple ontologies and the ability of selecting the set of ontologies (or ontology excerpts) to be applied to each given problem;
- the introduction of hand-made ontologies that may take into advantage users particular knowledge of the problem, correcting local problems;
- the comprehensive annotation of full-text documents that may support further relation extraction as well as more immediate analysis;
- the evaluation of available *Biomedical Part-of-Speech* (BPOS) taggers to help in the relation and interaction extraction.

Apart from plain observation of document annotations, users benefit from ontology support and document summaries. Each annotation identifies the category of the term as well as its entry in the ontology, providing both general and detailed information. Document summaries list the annotated terms, their

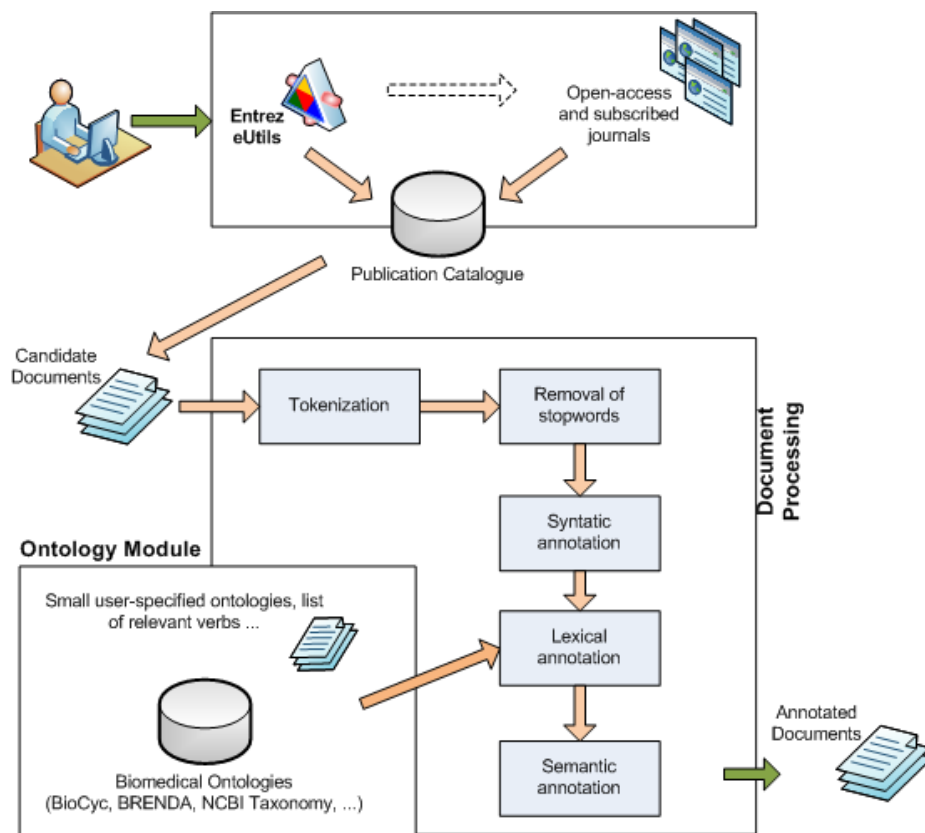


Fig. 1. General architecture of SysBio Explorer.

classes and frequency of annotation. These data may be used in the assessment of document relevance as well as to construct authoring networks. The identification of the documents that address a particular problem (e.g. a reaction or a pathway) may limit RE to such documents, improving the relevance of the acquired information. Likewise, the identification of the main research areas of each given author or team or the researchers that are working on a particular subject may be interesting in terms of IR and collaboration.

3.1 Document Retrieval

Commonly, *Biomedical Information Retrieval* (BIR) is based on abstract keyword matching, because there is a fair number of bibliographic databases that may support such task and most full-text documents require journal subscription. Two problems arise from this decision: the evaluation of document relevance is based on a small, general part of the document and such part cannot provide detailed information (e.g. regulatory and metabolic data).

In order to account for such problems, SysBio uses a two-stage BIR approach: the initial search is based on abstract contents, but, whenever possible, full texts are retrieved; then, NER procedures are used to unveil document contents and sustain further relevance assessment. The framework uses MEDLINE's PubMed bibliographic search facility, which is free of charge and supports external "calls" through Entrez Programming Utilities (eUtils). Specifically, it employs the `Bio::Biblio` package of BioPerl¹ to perform the keyword-based abstract searches and the `WWW::Mechanize`² package to automate the interaction with open-access and subscribed journals. Problem related information (user-specified keywords), publication details, abstracts and full-texts are recorded in SysBio's catalogue in order to perform BTM and support later cross-reference of TM results as well as case study analysis.

3.2 Ontology Definition and Integration

The ontology module aims at providing the means to integrate available ontologies as well as to create small, problem-specific ontologies. The package `Biblio::Thesaurus`³ [13] is used for managing the overall ISO monolingual and multilingual thesaurus[1,2] specification. It supports:

- **Mathematical Properties** — relation properties like inversion, symmetry, transitivity and reflexivity will make the ontology auto-completion active, making it easier to maintain ontology coherence;
- **Range and Domains** — differentiate between inter-term relations and external relations. Relations like scope notes, URLs or bibliographic links can provide additional information;
- **Multi-lingue Entries** — ontologies can include term definitions in more than one language.
- **Transitive Closure** — given a set of relations, `Biblio::Thesaurus` is able to compute the transitive closure for any specific term, making it easy to extract sub-ontologies regarding some specific knowledge area.

So far, the framework has injection functions for the BioCyc data bank, the NCBI Taxonomy and the BRENDA's enzyme ontology[3,10] and the inclusion of injection functions for UniProt, Gene Ontology (GO) and GeneBank resources are planned in short-term. Small, hand-made ontologies and a domain-specific list of verbs may be used in particular problems in order to perform disambiguation or to provide additional information. Table (1) lists some of the current relations.

The module allows the definition of the active parts of the ontology, i.e., to narrow down the ontology to be used by the NER module only to terminology

¹ <http://www.bioperl.org/>

² <http://search.cpan.org/dist/WWW-Mechanize/>

³ Although the Perl module is named `Biblio::Thesaurus` it uses a broad definition of thesaurus, where relations are user-defined, thus very close to the standard definition of ontology.

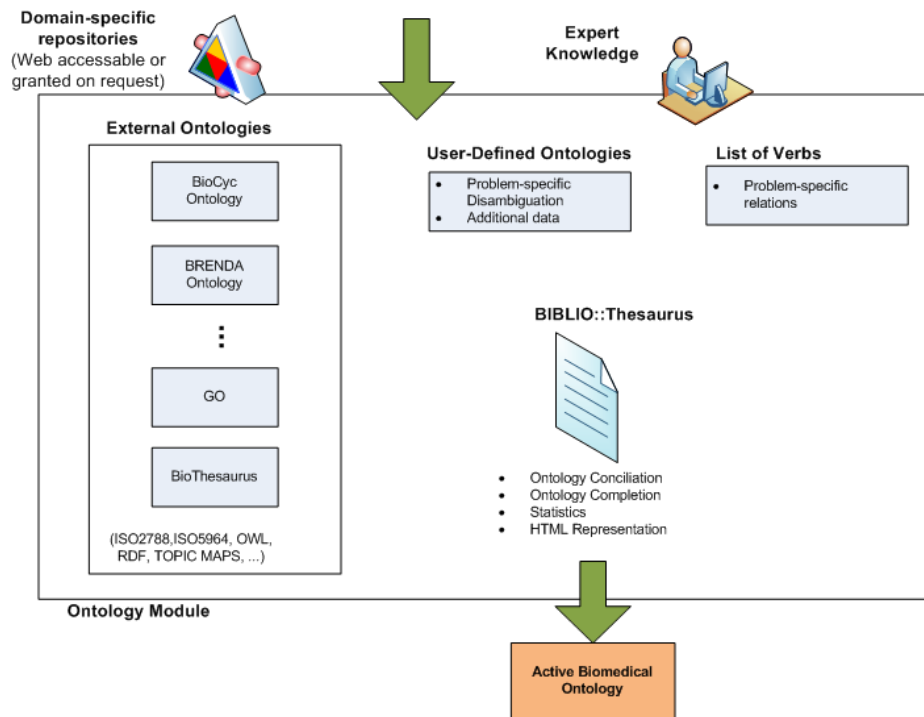


Fig. 2. SysBio Ontology Module.

related to the problem in question. For example, the sub-ontology may concern a given organism or set of organisms or may include only enzymatic information. It all depends on the particular problem.

Table 1. Subset of the used relations

Relation	Symmetric	Semantic
IOF	INST	A is a instance of B
POF	HAS	A is a part of B
SYN		A is a non-preferential term of B (e.g. systematic name, recommended name, ...)
DOM		Generic category of A
DIV		Taxonomy division
TYPE		Type of A (domain-oriented)
SN	—	Scope note relates a term with a brief description of the intended usage of the term/concept

3.3 Document Processing and NER

Document processing involves PDF file conversion, conventional text processing and biomedical-specific text processing. The conversion of PDF files into plain ASCII files is based on pdftotxt open-source tool⁴. Text is tokenized, common English stopwords are ignored, and the GENIA BPOS tool⁵, performs domain-specific linguistic POS annotation⁶.

SysBio's NER module is based on the ontology produced by Biblio::Thesaurus and aims at the configurable recognition, normalisation and classification of relevant terms. A configuration file allows the user to specify the biomedical entities that should be annotated and the ontologies and user-made specifications (ontologies, lists and rules) that should sustain the process while a Cascade Style Sheet (CSS) file specifies the contents and visual effect of such annotation. On the other hand, a term rewriting system, i.e., a reduction system in which rewrite rules apply to terms, encompasses the set of active annotation rules. The system was implemented using the `Text::RewritingRules` package and has strong pattern matching skills that allow the specification of several kinds of rules, from simple substitution rules to conditional and evaluated rules:

```
left hand side ==> right hand side
left hand side =e=> right hand side
left hand side ==> right hand !! condition
```

Furthermore, it provides the means to manage the rules swiftly without altering the rest of the annotation module. The initial set of rules was defined after exploring term patterns in the ontology terminology. Presently, rewriting rules target single-word terms and hepta, hexa, penta, tetra, tri and bi-grams, evaluating the corresponding class counts. New problems may demand the adaptation of such rules or the inclusion of new (general or problem-specific rules).

The NER module delivers a XML file for each annotated document and a general statistics file for the set of processed documents. Entities are differentiated by colours that identify term classes and each annotation links to the corresponding ontology information which in turn allows the access to external repositories (e.g. through GO codes). Apart from actual annotations, each XML file provides a summary of the annotated terms and their corresponding weighted occurrences.

$$author \rightarrow (term \rightarrow occurrences)$$

Such summaries support further RE and the assessment of document similarity aiming at both document relevance assessment and the construction of authoring networks.

⁴ <http://www.foolabs.com/xpdf/>

⁵ <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/>

⁶ Currently we are not taking advantage of the POS tagging but it is already included in the processing pipeline as it would be of use for relation and interaction extraction.

3.4 Document Similarity Analysis

Similarity analysis addresses:

- the refinement the list of candidate documents retrieved from PubMed;
- the identification of the biomedical entities that an author refers the most, generating an authoring list;
- the cross-reference of authoring lists aiming at identifying researchers working on particular research domains;
- the cross-reference of authoring lists aiming at identifying researchers with similar research domains.

The cosine similarity, a common measure in IR, sustains this line of analysis. Assuming that document vectors with many words in common will point in roughly the same direction and thus, the angle between two document vectors is a good measure of their similarity.

4 Experiments

Experiments addressed problems concerning the well-known organisms *E. coli*, *S. cerevisiae* and *H. pylori*. The results of SysBio BIR module are listed in Table 2, indicating the number of documents that matched the posted queries, the number of available abstracts and the number of retrieved full-text documents.

Table 2. General statistics about case studies.

Case Study	PubMed results	Abstracts	PDFs
<i>E.coli</i> stringent response	294	286	105
<i>H.pylori</i> virulence factors	399	388	98
<i>S.cerevisiae</i> ethanol production	660	658	136

The NER module recognised biological entities and their corresponding main classes (Table 4 and Table 3) and biologically related verbs. So far, the configurable annotation scheme (colour and XML tags) targets organisms, genes, proteins, reactions, compounds and RNA class members, and biologically relevant, user-specified verbs. Additionally, when a given term matches an annotation rule, but there is no information about the corresponding class, a default tagging is used.

The comparison of full-text and abstracts results provides additional insights about information coverage and density. Usually, abstracts contain a best ratio of keywords per total of words, but other article sections (such as introduction, methods, results, and discussion sections) may be a better source of biologically relevant data, namely metabolic and regulatory data [12,11].

Table 3. Comparison of class annotation in full texts versus abstracts.

	E. coli		H. pylori		S. cerevisiae	
	Full	Abs.	Full	Abs.	Full	Abs.
organism	12.23	13.53	30.07	35.04	11.89	13.27
compounds	21.87	25.46	14.76	12.39	40.38	47.54
genes	44.37	38.94	37.82	30.42	21.04	12.39
proteins	15.34	17.10	11.43	12.86	6.67	5.16
reactions	1.49	1.73	5.16	8.66	3.96	5.12
pathways	0.31	0.05	0.04		1.51	1.62
unknown	1.81	1.85	0.83	0.96	14.15	14.70

Table 4 and Table 5 present top annotation results and mid-list results for full texts and abstracts, respectively. For a given term t_i its relative frequency on document d is

$$p_i^d = \frac{\text{occurrences}(t_i, d)}{\sum_{j \in T} \text{occurrences}(t_j, d)}$$

where T is the multiset of annotated entities in document d . For each problem, the mean of relative term frequencies over the set of documents was calculated.

Table 4. Top 15 relative frequencies of terms in full-text documents.

A	B	C
ppGpp 9.21	Helicobacter pylori 10.89	ethanol 12.81
relA 4.74	CagA 5.51	yeast 11.09
Escherichia coli 3.79	cagA 4.68	Saccharomyces 6.02
spoT 2.19	vacA 4.18	glycerol 2.83
...
fis 1.17	Med 1.02	Ethanol 1.21
Proc 1.00	Proc 1.01	xylitol 1.11
GTP 0.95	der 0.96	CO2 1.06
SpoT 0.95	vacuolating cytotoxin 0.96	lactate 0.84
ATP 0.92	iceA 0.91	XDH 0.74
guanosine 0.89	cagE 0.86	Fermentation 0.73

It is possible to observe that top annotations refer to general terms

5 Conclusions and Future Work

BTM is delivering important breakthroughs in terms of automatic literature curation. Yet, most works focus on scientific contribution and neglect real-world application. Even though techniques are of major importance, the biomedical community has to acknowledge BTM contribution to the resolution of its current problems as well as to the evolving of its analysis abilities. In this regard, the

Table 5. Top 15 relative frequencies of terms in abstracts.

E. coli		H. pylori		S. cerevisiae	
ppGpp	13.14	Helicobacter pylori	11.10	ethanol	23.22
relA	6.71	cagA	8.64	yeast	10.50
RNA	6.51	CagA	7.53	Saccharomyces	8.16
Escherichia coli	5.64	VacA	6.98	glycerol	3.42
guanosine	2.44	vacA	6.80	yeasts	2.06
...
rel	1.52	babA2	1.81	acetaldehyde	0.83
mRNA	1.25	iceA	1.46	xylitol	0.73
synthetase	1.14	Helicobacter	1.37	CO2	0.67
GTP	1.12	oipA	0.96	trehalose	0.64
SpoT	1.09	gerbils	0.75	alcohol dehydrogenase	0.63

SysBio Explorer presents the following contributions: full-text retrieval and processing in order to extract detailed information, namely regulatory and metabolic data; the definition of injection functions for prominent biological repositories and the construction of domain-specific ontologies; biomedical ontology-based entity named recognition and annotation; user-friendly annotation; and the construction of authoring networks.

The real-world problems meant to demonstrate the usability and usefulness of the framework without disregarding present limitations, both in terms of ontology management, text processing and analysis.

Besides ensuring the integration of additional information resources, such as UniProt, GeneBank or KEGG, future work targets the following problems:

- the recognition of term variants,
- the resolution of term abbreviations and
- the extraction of relevant relations, namely regulatory and metabolic relations, combining BPOS tagging with state-of-the-art BTM.

References

1. *ISO 2788 – Guidelines for the establishment & development of monolingual thesauri*. International Organization for Standardization.
2. *ISO 5964 – Guidelines for the establishment & development of multilingual thesauri*. International Organization for Standardization.
3. J. Barthelmes, C. Ebeling, A. Chang, I. Schomburg, and D. Schomburg. Brenda, amenda and frenda: the enzyme information system in 2007. *Nucleic Acids Res.*, 35(Database issue):D511–D514, 2007.
4. N. Dimililer and E. Varoglu. Recognizing biomedical named entities using svms: Improving recognition performance with a minimal set of features. In *Knowledge Discovery in Life Science Literature*, pages 53–67. 2006.
5. Z. Z. Hu, M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker, and C. H. Wu. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, 21(11):2759–2765, 2005.

6. J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. Genia corpus—semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):i180–i182, 2003.
7. Z. Kou, W. W. Cohen, and R. F. Murphy. High-recall protein entity recognition using a dictionary. *Bioinformatics*, 21(Suppl 1):i266–i273, 2005.
8. H. F. Liu, Z. Z. Hu, M. Torii, C. Wu, and C. Friedman. ”quantitative assessment of dictionary-based protein named entity tagging. *Journal of the American Medical Informatics Association*, 13(5):497–507, 2006.
9. J. Natarajan, D. Berrar, C. J. Hack, and W. Dublitzky. Knowledge discovery in biology and biotechnology texts: A review of techniques, evaluation strategies, and applications. *Critical Reviews in Biotechnology*, 25(1-2):31–52, 2005.
10. I. Schomburg, A. J. Chang, O. Hofmann, C. Ebeling, F. Ehrentreich, and D. Schomburg. Brenda: a resource for enzyme data and metabolic information. *Trends in Biochemical Sciences*, 27(1):54–56, 2002.
11. M. J. Schuemie, M. Weeber, B. J. A. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jelier, B. Mons, and J. A. Kors. Distribution of information in biomedical abstracts and full-text publications. *BMC Bioinformatics*, 20(16):2597–2604, 2004.
12. P. K. Shah, C. Perez-Iratxeta, P. Bork, and M. A. Andrade. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4, 2003.
13. A. M. Simões and J. J. Almeida. Library::* — a toolkit for digital libraries. In *ELPub 2002 - Technology Interactions*, 2002.
14. C. J. Sun, Y. Guan, X. L. Wang, and L. Lin. Biomedical named entities recognition using conditional random fields model. In *Fuzzy Systems and Knowledge Discovery*, pages 1279–1288, 2006.
15. R. T. 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*, 7(5):S11, 2006.
16. T. H. Tsai, W. C. Chou, S. H. Wu, T. Y. Sung, J. Hsiang, and W. L. Hsu. Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. *Expert Systems with Applications*, 30(1):117–128, 2006.