An Ontology-Based Approach To Systems Biology Literature Retrieval and Processing

Anália Lourenço[1,2], Alberto Simões[1], José João Almeida[1], Miguel Rocha[1], Isabel Rocha[2], and Eugénio C. Ferreira[2]

[1] Departament of Informatics / CCTC - University of Minho
Campus de Gualtar, 4710-057 Braga - PORTUGAL
{analia,ambs,jj,mrocha}@di.uminho.pt
[2] IBB - Institute for Biotechnology and Bioengineering
Center of Biological Engineering - University of Minho
Campus de Gualtar, 4710-057 Braga - PORTUGAL
{irocha,ecferreira}@deb.uminho.pt

Abstract. This paper details the *SysBio Explorer*, a Systems Biology Literature Retrieval and Processing Framework, whose aim relies on the automatic inference of regulatory and metabolic networks based on biomedical literature. The *SysBio Explorer* does not focus on any organism or problem in particular and encompasses a number of 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 authoring networks are also within its present scope. The framework is described in terms of its design requirements and implementation decisions, exposing current achievements, but also highlighting present obstacles and future work. 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, 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 inference of 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 deal with 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 search for 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 verbs related to biological relationships, 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 will also be pursued in the future.

The paper details experiments over real-world problems concerning the organisms *E.coli*, *S.cerevisiae* and *H.pylori*. Such experiments aim at demonstrating the usefulness of the framework 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.

Disregarding the approach that is chosen for a particular problem, knowledge acquisition is always necessary and poses challenging problems. The manual curation of a representative number of documents for a particular problem is time-consuming. On the other hand, annotated corpora (e.g. the GENIA corpus [6]) are built from the results of particular keyword-based queries and thus are biased to a particular domain/problem and cannot provide decision models that perform well in other user-specified problems. Such biased annotation resources are suitable for technique benchmarking, but cannot be used on general, user-specified problem annotation.

Available biological repositories may provide general information resources, but often multi-repository integration is scarce and the quality of the contents is somewhat disputable. 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.

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 applicability. 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 inference of the corresponding networks. Besides supporting state-of-the-art BTM, this framework differs from existing work because it is meant for common use by Biology researchers without specific tutoring. It provides the means to take into account as much information resources as available, detaching from particular organisms or problems. Furthermore, it tackles full-text documents 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;

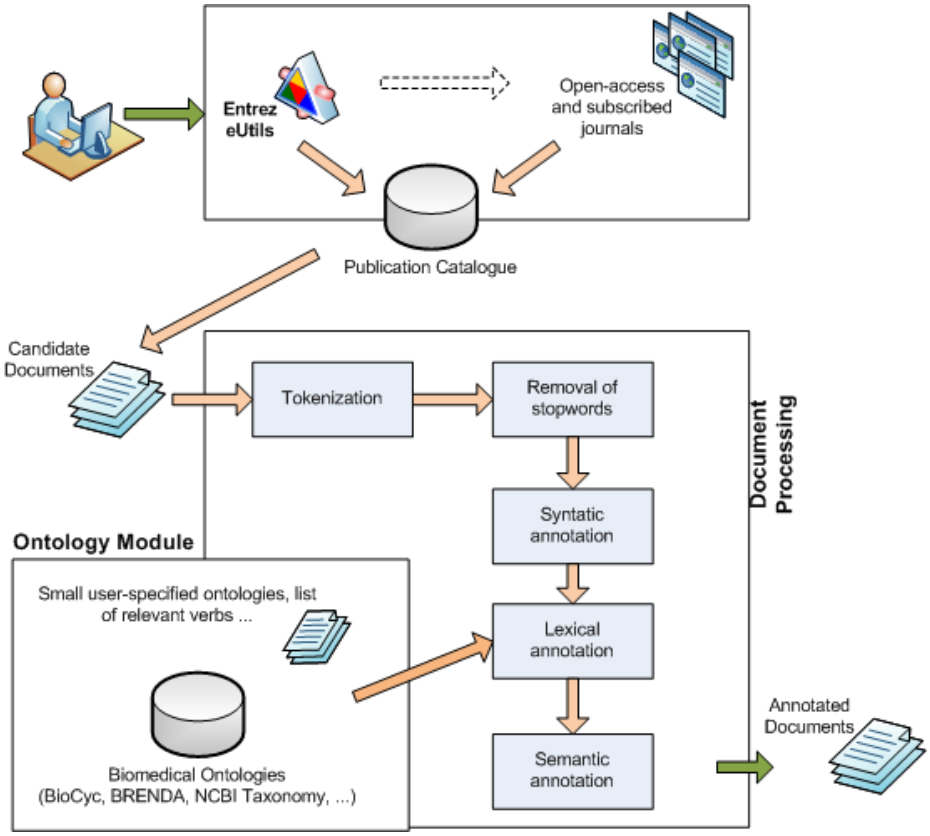

Fig. 1. General architecture of SysBio Explorer.

- 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 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 restrict 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 BioPerl[3] to perform the keyword-based abstract searches and the WWW::Mechanize[4] 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`[5] [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

[3] `http://www.bioperl.org/`

[4] `http://search.cpan.org/dist/WWW-Mechanize/`

[5] 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.

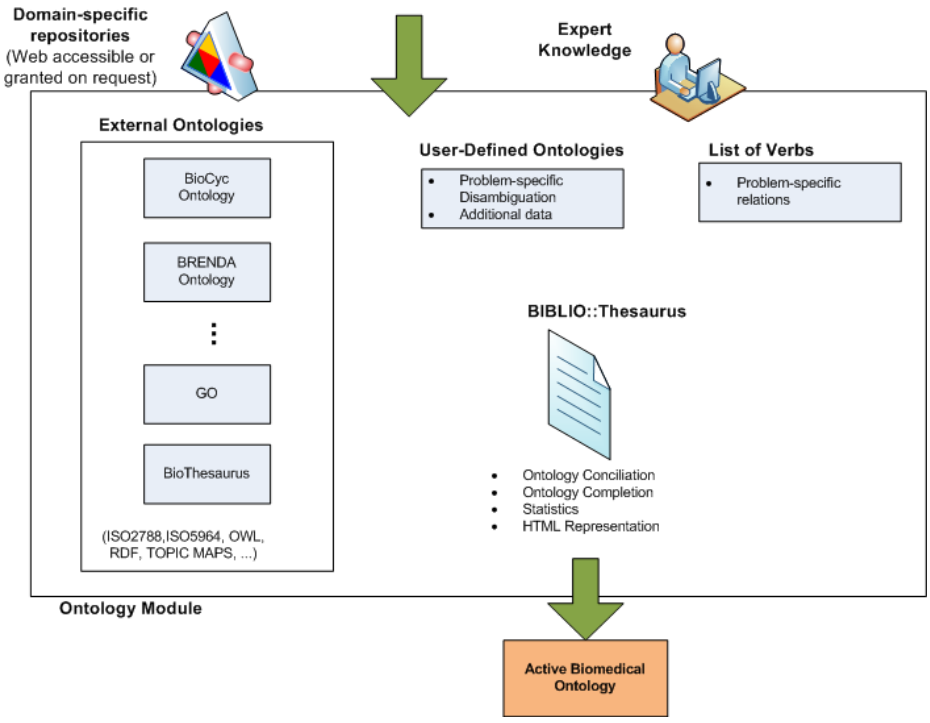

Fig. 2. SysBio Ontology Module.

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.

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

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 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.

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 tool[6]. Text is tokenized, common English stopwords are ignored, and the GENIA BPOS tool [7] performs domain-specific linguistic POS annotation[8].

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 will 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::RewrittingRules` 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. Such summaries support further RE and the assessment of document similarity aiming at both

[6] `http://www.foolabs.com/xpdf/`

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

[8] 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.

document relevance assessment and the construction of authoring networks. So far, SysBio does not yet support RE procedures, but it already supports document similarity analysis, based on cosine similarity, a common measure in IR. Similarity analysis addresses:

- the refinement of 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.

4 Experiments

Experiments addressed problems concerning three specific queries related to 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
E.coli stringent response	294	286	105
H.pylori virulence factors	399	388	98
S.cerevisiae ethanol production	660	658	136

So far, the NER module identifies organisms, genes, proteins, reactions, compounds and RNA class members, and biologically relevant, user-specified verbs. Additionally, the configurable annotation scheme (colour and XML tags) supported the annotation of unclassified terms, i. e., when a given term matches an annotation rule, but there is no information about the corresponding class, a default tagging is used.

Document summaries and general statistics present relative frequency of term annotations. For a given term t_i its relative frequency on document d is

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

where T is the multiset of annotated entities in document d and $occurrences(t_i, d)$ stands for the number of occurrences of t_i in document d. For each problem, the mean of the relative term frequencies over the set of documents was calculated.

Table 3. Class annotation in full texts.

	E. coli	***H. pylori***	***S. cerevisiae***
organism	12.23	30.07	11.89
compounds	21.87	14.76	40.38
genes	44.37	37.82	21.04
proteins	15.34	11.43	6.67
reactions	1.49	5.16	3.96
pathways	0.31	0.04	1.51
unknown	1.81	0.83	14.15

Table 3 presents the relative frequency of annotation of each class and Table 4 details annotation frequencies for the most common terms. Problem-specific analysis of the annotations required the assistance of specialists. However, some of the results were fairly easy to interpret based on the general context of the problems. Ethanol production in *S. cerevisiae* is a metabolic process and *E. coli* stringent response and the virulence factors in *H. pylori* are related to regulatory processes. In this sense, the ranks of annotated classes make perfect sense: compounds is the most annotated class for *S. cerevisiae*, while genes is the most annotated class for *E. coli* and *H. pylori*. Additional biologically-related stopwords, i.e., biological terms that are also common English terms and can easily raise the rate of false positive annotations were also identified. For example, terms such as *Proc*, *Appl* or *Med* should be ignored unless the user specifies otherwise.

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

E. coli		***H. pylori***		***S. cerevisiae***	
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
mRNA	1.97	VacA	3.83	Appl	2.72
tRNA	1.90	cag	3.41	yeasts	1.84
rRNA	1.79	human	2.33	ATP	1.47
RelA	1.61	urease	1.51	NADH	1.40
synthetase	1.44	Helicobacter	1.48	acetate	1.32
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

4.1 Abstracts versus Full-Text Contents

The systematic comparison of full-text and abstracts results may provide 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].

It was out of the scope of this paper to perform such analysis. Yet, it was considered interesting to analyse an example. The specialists chose a document related to *E. coli* and concerning stringent response, specifically the activity of inhibitors in transcription. They described the problem as regulatory-related, i.e., a problem that would involve many references to genes in its debate. More important, they stated that there were a number of particular entities whose relation to the problem is well-known and may be used both to certify document's relevance and to discover previously unknown information. For example, genes such as *relA* and metabolites such as *ppGpp* are known to be important in this particular problem, but their presence along with the identification of other entities may highlight the particular role of other entities and allow the establishment of relations among them.

SysBio's results (Table 5) showed that abstract annotations are mostly general problem-related terms, identifying only two genes (*relA* and *grpE*) and one enzyme ($leucyl - tRNAsynthetase$). On the other hand, full-text annotation highlights other well-known, problem-related entities (e.g. *ppGpp* and $isoleucyl - tRNAsynthetase$), but also many other entities (over than 100 different gene annotations) that provide more detailed (and eventually new) knowledge (e.g. references to pathways and reactions or mentions to other organisms).

5 Conclusions and Future Work

BTM is delivering important breakthroughs in terms of automatic literature curation. Yet, most work 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 *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 usefulness of the framework, 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,

Table 5. An example-oriented comparison of full-text and abstract content coverage and diversity.

Term	# Abs. Annotations	# Full. Annotations
relA	2	41
Escherichia coli	1	23
leucyl-tRNA synthetase	1	2
mRNA	1	16
grpE	1	12
trna	0	20
ilvC	0	18
aroF	0	17
metE	0	15
Salmonella	0	13
ppGpp	0	6
acivicin	0	5
isoleucyl-tRNA synthetase	0	4
histidine biosynthesis	0	1
glyoxylate bypass	0	1

– 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.

Acknowledgments

The work of Anália Lourenço is funded by the research project *recSysBio - A Systems Biology approach for optimization of recombinant fermentation processes* (ref. POCI/BIO/60139/2004) of the *IBB - Institute for Biotechnology and Bioengineering* of the University of Minho, financed by the *Fundação para a Ciência e Tecnologia*.
Alberto Simões has a scholarship from the *Fundação para a Computação Científica Nacional* and the work reported here has been partially funded by the *Fundação para a Ciência e Tecnologia* through project POSI/PLP/43931/2001, co-financed by POSI, and by POSC project POSC/339/1.3/C/NAC.

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.