BioLexicon: A Lexical Resource for the Biology Domain

Yutaka Sasaki [1] Simonetta Montemagni [3] Piotr Pezik [4] Dietrich Rebholz-Schuhmann [4]
John McNaught [1, 2] Sophia Ananiadou [1, 2]

[1] School of Computer Science, University of Manchester
[2] National Centre for Text Mining
MIB, 131 Princess Street, Manchester, M1 7DN, United Kingdom
Yutaka.Sasaki@manchester.ac.uk

[3] Istituto di Linguistica Computazionale, CNR, Via Moruzzi 1, 56124 Pisa, Italy

[4] EBI, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom

Abstract

Natural language processing technologies have advanced remarkably in the past two decades. However, biological terminology is a frequent cause of analysis errors when processing literature written in the biology domain. The BOOTStrep BioLexicon is a linguistic resource tailored for the domain to cope with these problems. It contains the following types of entries: (1) a set of terminological verbs; (2) a set of derived forms of the terminological verbs; (3) general English words frequently used in the biology domain; (4) domain terms. This comprehensive coverage of biological terms makes the lexicon a unique linguistic resource within the domain. This paper focuses on the linguistic aspects of the lexicon.

1 Introduction

Over the past twenty years, there have been remarkable advances in natural language processing (NLP) and text mining (TM) technologies. Various practical NLP/TM tools, such as part-of-speech taggers, chunkers, syntactic parsers and named entity recognizers, are now widely available.

However, text in biology exhibits different characteristics from general language documents such as newspaper articles. The biology domain demonstrates strong demands for the results of NLP/TM. However, it is also one of the most challenging domains for text processing (Ananiadou and McNaught, 2006).

Lack of coverage of the following types of terminological information makes NLP/TM tasks in this domain difficult:

- Large-scale domain-specific terminologies
- Domain-specific word usage
- Domain-specific relations between words

Technical terms are a major barrier to bio-text processing. A huge number of biological, chemical and medical terms appear in the literature and new terms are coined every day. Furthermore, there are many spelling and semantic variants of these terms representing the same biomedical entities in different written forms. For example, the BioThesaurus [1] contains more than 15 million gene/protein names, but still it does not cover the wide variety of variants of gene/protein names actually appearing in the literature.

Word usage can be idiosyncratic to the bio-domain as well. For example, *express* often indicates a specific biological process, *gene expression*, and takes as arguments specific types of named entities, such as gene and protein names.

In addition, there are many cases where words are related in a biology-specific manner. For example, the verb *retroregulate* has *retroregulation* as its nominal form and *retroregulatory* as its adjectival form. This extent

[1] http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml

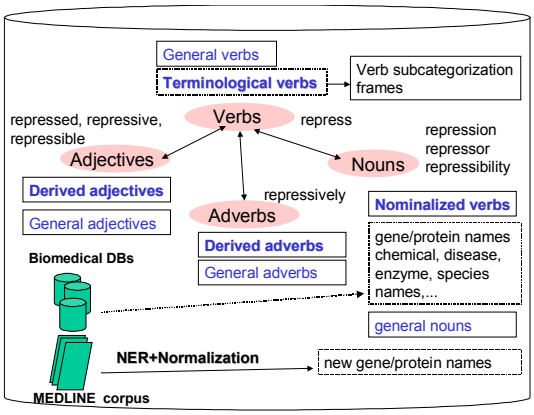

Figure 1 Overview of the Lexicon

of derivational relations between words in the biological domain cannot be fully covered by general English dictionaries and thesauri, *e.g.*, WordNet. To the best of our knowledge, there is no biology-specific lexicon that addresses the above linguistic issues.

2 Overview of the BioLexicon

Figure 1 shows an overview of the BioLexicon. It consists of four part-of-speech categories: verb, noun, adjective, and adverb. Each category accommodates terminological words and general language words. Biology terms, *e.g.*, gene/protein names, are either gathered from existing databases or automatically extracted from text. Other terminological words and their relations are manually curated. Inflections of general words are manually curated based on the MedPost dictionary (Smith *et al.,* 2004).

The database model of the lexicon follows the Lexical Markup Framework (LMF) (Francopoulo *et al*., 2006). The details of the database model were reported in Quochi *et al*. (2008).

3 Biology-relevant terminologies

The terminologies in the lexicon are fivefold: (1) verbs, (2) adjectives, (3) adverbs; (4) terminological nouns, and (5) biomedical terms. (1) – (4) have been manually curated.

(1) Terminological verbs
759 base forms (4,556 inflections) of terminological verbs.

(2) Terminological adjectives
1,258 terminological adjectives.

(3) Terminological adverbs
130 terminological adverbs.

(4) Nominalized verbs
1,771 nominalized verbs.

(5) Biomedical terms
Currently, the BioLexicon contains biomedical terms in the categories of cell (842 entries, 1,400 variants), chemicals (19,637 entries, 106,302 variants), enzymes (4,016 entries, 11,674 variants), diseases (19,457 entries, 33,161 variants), genes and proteins (1,640,608 entries, 3,048,920 variants), gene ontology concepts (25,219 entries, 81,642 variants), molecular role concepts (8,850 entries, 60,408 variants), operons (2,672 entries, 3,145 variants), protein complexes (2,104 entries, 2,647 variants), protein domains (16,940 entries, 33,880 variants), Sequence ontology concepts (1,431 entries, 2,326 variants), species (482,992 entries, 669,481 variants), and transcription factors (160 entries, 795 variants).

In addition to the existing gene/protein names, 70,105 variants of gene/protein names have been newly extracted from 15 million MEDLINE abstracts. Section 5 describes the methods used.

3.1 Terminological verbs

Terminological verbs have been manually curated through examination of biomedical literature. As a result, 759 verbs were selected.

Following the selection of verbs, three types of orthographic variants were added to the lexicon.

- British/American spelling variants
e.g., *acetylise* (British)/*acetylize* (American) or *harbour* (British)/*harbor* (American)

- Hyphenation variants
e.g., *co-activate* and *coactivate*

- Combination of the above two
e.g., co-localise (British), colocalise (British), co-localize (American), colocalize (American)

Inflectional forms are all enumerated in our lexicon. The following verbal inflections have been completely curated.

VV base form
VVD past tense
VVN past participle
VVZ third person singular present
VVG gerund or present participle

The above parts-of-speech follow the Penn Treebank POS tags (Santorini, 1990).

3.2 Derived forms of terminological verbs

Our strategy was to expand the terminology from terminological verbs to derived forms. Three types of derivational relations of the terminological verbs have been introduced. Frequently, nominalized verbs play the same role as verbs. Adjectival and adverbial derived forms may also be used to represent biological events and processes in the same context as their associated verbs. For text mining applications, it is important to cover these possibilities as far as those derivations are linguistically correct.

(1) Nominalization

Nominalized verbs are verbs that are used as nouns. A verb can be nominalized with or without morphological transformation. For example, the nominalized forms of *regulate* are *regulation* and *regulator*. Following Comrie and Thompson (2007), we identified two kinds of nominalization.

(i) Action/state nouns

The noun expresses an action or state of the verb from which it is derived, *e.g.*,

act (v) → action (n),
act (v) → act (n),
act (v) → acting (n).

(ii) Agentive nouns

The noun has an 'agent' role to the verb from which it is derived, *e.g.*,

act (v) → actor (n)

(2) Adjectival derivation

The derivational relation between adjectives and the verbs from which they are derived was manually curated, because there is no dictionary that fully covers adjectival derivations of biological terms. *E.g.*,

act (v) → actable (adj.),
act (v) → active (adj.).

(3) Adverbial derivation

The derivational relation between adverbs and the verbs from which they are derived were also manually curated, *e.g.*,

act (v) → actively (adv.)

3.3 Biomedical terms

Existing biological databases have served as the first source of many nominal types of terms represented in the BioLexicon. Detailed information can be found on the BOOTstrep web site. (Bootstrep, 2008). Such resources are characterized by a high coverage of biological entities and they contain terms annotated with widely recognized and interoperable accession number (*e.g.,* UniProt). On the other hand, some terms imported from existing resources are assigned to concept identifiers in the process of automatic curation. Moreover, although biological ontologies and controlled vocabularies are meant to represent a wide range of concepts, they are not designed to reflect the exact wording found in the scientific literature. Therefore, some initial filtering of potential terms was necessary before they could be included in the BioLexicon. As an example, terms of proteins identified in the course of high-throughput experiments such as *hypothetical protein* were ignored due to their low information value. Also, a small number of highly ambiguous terms such as generic enzyme names were manually annotated as such. Other indications of a term's discriminatory power available in the BioLexicon include its frequency in Medline and the British National Corpus, as they have proven useful in the task of named entity recognition (Pezik *et al*., 2008).

The choice of these types of terms can be explained in two ways. Firstly, we felt it necessary to include the most common semantic types relevant to the biology domain, such as terms denoting gene and protein names, as well as terms for chemicals of biological interest or species

names. Secondly, including the smaller and more focused sets for terms such as operon names or sequence ontology terms was motivated by the intention to provide links from the BioLexicon to the Gene Regulation Ontology (Beisswanger *et al*., 2008) and make it suitable for text mining applications dealing with gene regulation topics.

4 General language words

To cover general language words that are used in biology, we have adopted words from the MedPost dictionary. This is distributed as a part of the MedPost POS tagger package and is available copyright free.[1] The dictionary consists of words appearing in MEDLINE abstracts.

The following numbers of entries were generated.

- 496 verbs (2,976 inflectional forms)
- 2,316 adjectives (2,385 inflectional forms)
- 428 adverbs (440 inflectional forms)
- 5,012 nouns (6,182 inflectional forms)

Inflections produced for verbs from the MedPost dictionary are the same as for terminological verbs. The POS types NN and NNS were assigned to the singular and plural forms of nouns, respectively.

Comparative and superlative forms of adjectives and adverbs were completed on the basis of the MedPost dictionary entries.

Since that dictionary was created for the purposes of a statistical POS tagger for the biomedical domain, it is incomplete from a linguistic point of view. For example, *common* and *commonest* are accommodated by the dictionary; however, *commoner* is not. Therefore, inflections of words in the dictionary were manually curated and added to the BioLexicon.

5 Biological term variants extracted from text

In addition to biomedical terms gathered from existing databases, the lexicon accommodates new variants of gene/protein names extracted from text.

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

Table 1 NER performance

		R	P	F
Sequential labeling	Full	79.85	68.58	73.78
	Left	84.82	72.85	78.38
	Right	86.60	74.37	80.02

The extraction process consists of two steps. The first step identifies gene/protein names in text. Then, the second step maps new variants to existing entries.

This section provides a brief summary of the named entity recognition (NER) and term normalization used to populate the lexicon with gene/protein names extracted from biomedical literature.

5.1 Named Entity Recognition

For NER, we used our dictionary-based statistical named entity recognition tool (Sasaki *et al*., 2008).

The tool was trained with Conditional Random Fields (CRFs) (Lafferty *et al*., 2001) on the JNLPBA-2004 training data (Kim, 2004) and the Genia corpus (version 3.02) (Kim *et al*., 2003).

The test data used is the JNLPBA-2004 test set, which is a set of tokenized sentences extracted from 404 separately collected MEDLINE abstracts, where the term class labels were manually assigned, in accordance with the annotation specification of the Genia corpus.

Following the data format of the JNLPBA-2004 training set, our training and test data use the IOB2 labels, which are "B-protein" for the first token of the target sequence, "I-protein" for each remaining token in the target sequence, and "O" for other tokens. The window size was set to ±2 tokens of the current token.

Table 1 shows the evaluation results. Results are expressed according to recall (R), precision (P), and F-measure (F), which here measure how accurately the various experiments determine the left boundary (Left), the right boundary (Right), and both boundaries (Full) of protein names. The The F-score of the model trained with all the features was 73.78, which is the second best score for protein name recognition among research reported using the standard JNLPBA-2004 data set.

Gene/protein names identified by CRF classifiers with a probability greater than 99% are

selected as new gene/protein variant candidates from 15 million MEDLINE abstracts.

5.2 Term mapping

Terms automatically extracted from text were mapped to existing gene/protein name entries, which are given standard semantic identifiers called UniProt Accession Numbers. For efficiency reasons, term mapping was conducted through term normalization. Since the lexicon contains about two million gene/protein names, straightforward similarity calculation of term pairs is not practical: when an NER component extracts tens of millions of gene/protein name candidates from a corpus, the similarity distance of $2 \cdot 10^{13}$ pairs of terms must be calculated. This amount of computation can be drastically reduced to 10^7 normalizations and index lookups.

The normalization steps are as follows:

1. Create an inverse index that maps normalized forms to UniProt Accession Numbers.
2. Normalize newly extracted terms.
3. Lookup the inverse index to find UniProt Accession Numbers of the new terms.

There are several ways to normalize biomedical terms. We employed a method (Tsuruoka *et al*., 2007) where the normalization rules were automatically generated from a dictionary in which terms are clustered according to UniProt Accession Numbers. A brief summary of the method is as follows:

The method finds string-rewriting rules one by one based on the following complexity measure:

$$(\text{complexity}) = (\text{ambiguity}) \times (\text{variability})^{\alpha}$$

where the ambiguity quantifies how ambiguous the terms are in the dictionary, the variability value quantifies how variable the terms are, and α is the constant that determines the trade-off between ambiguity and variability.

Finding string rewriting rules is quite straightforward. We can represent any pair of terms *x* and *y* as follows:

$x = LXR$
y = LYR

where *L* is the left common substring shared by strings *x* and *y*, *R* is the right common substring, and *X* and *Y* are the substrings in the center that are not shared by the two strings. From this representation, we create the rule that replaces *Y* with *X*, which will transform *y* into *x*.

According to the experimental results reported in Tsuruoka *et al*. (2007), normalization performance is the same as normalization rules hand-crafted by domain experts. We generated 1,000 normalization rules, using the gene/protein names gathered from existing databases as the dictionary for normalization rule generation.

Terms mapped to more than 10 accession numbers are considered too ambiguous and filtered out from the new variant list. As a result, 70,105 variants of gene/protein names were extracted from 15 million MEDLINE abstracts.

6 Biomedical usages

In the lexicon, terminological verbs are linked to verb subcategorization frames (SCFs) which were acquired through unsupervised automatic acquisition techniques from linguistically pre-processed domain corpora. In the biomedical field, there is a strongly-felt desideratum that subcategorisation patterns should include strongly selected modifiers (such as location, manner and timing), as these are deemed to be essential for the correct interpretation of texts (Tsai *et al*., 2007). According to this, we adopted a "discovery" approach to SCF acquisition based on a looser notion of SCFs, which include typical verb modifiers in addition to strongly selected arguments.

In order to meet this basic requirement, a deep level of syntactic annotation was selected as the starting point for SCF induction. For this purpose, we used the Enju syntactic parser for English (Miyao *et al*., 2003)[1], characterised by a wide-coverage probabilistic HPSG grammar and an efficient parsing algorithm, and whose output is returned in terms of predicate-argument relations. In particular, we used the Enju version adapted to biomedical texts (Hara *et al*., 2005).

The SCF induction process was performed through the following steps:

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/enju/

1. syntactic annotation of the acquisition corpus with Enju (v2.2). The acquisition corpus included both MEDLINE abstracts and full papers containing a total of approximately 6 million word tokens;
2. for each verbal occurrence, extraction of the observed dependency sets (ODSs). Note that the order of the dependencies in each ODS is normalised and does not reflect their order of occurrence in context;
3. induction of relevant SCF information associated with a given verb.

For each observed dependency set, the conditional probability given the verb type v was computed: thresholding was used, to filter out noisy frames (i.e., frames containing not only arguments and strongly selected modifiers, but also adjuncts) as well as possible errors of either parsing or ODS extraction. An ODS with an associated probability score beyond a certain threshold is selected as eligible SCF for that verb type.

Careful analysis of acquired SCFs revealed that many of the strongly selected modifiers were spread over different frames and that, even by lowering filtering thresholds, they either disappeared from the final output or their role was radically underestimated. We thus decided to complement acquired SCF information with information about individual dependencies of verbs. To detect typical verbal dependencies, corresponding to either arguments or strongly selected modifiers, we used the log likelihood score (henceforth ll (Dunning, 1993)). This is a logarithmic measure of the degree of correlation between v and each dependency type, gauged by comparing their joint probability with the probability of finding them together by chance, given their independent marginal distributions.

Due to the observed complementarity between acquired SCF and individual dependency information and its potential usage in different text mining applications, we decided to include both information types in the lexicon. SCF and dependency information was acquired for 759 orthographic variants of different terminological verbs, corresponding to 658 different base forms (see section 3.1). In particular, the lexicon includes 1,410 verb-SCF associations, involving 97 different SCF types, and 1,718 verb-dependency associations, involving 44 dependency types. For each SCF, the following information types are specified: its conditional probability given the verb, and the percentage of times it occurs with the verb in the passive voice. This latter information type is particularly useful to account for SCFs typically associated with the verb used in the passive voice: this is the case, for instance, of the verb *find* whose frame ARG1#ARG2#TO-INF# is typically (*i.e.*, 89% of the time) associated with passive contexts (*e.g., This was found to be interesting*). Concerning individual dependencies, the lexicon includes information about its association with respect to the verb, expressed in terms of the ll score, and – again – the percentage of times it occurs with the verb in the passive voice. Tables 2 and 3 show examples of subcategorization information stored in the lexicon for the verb *acquire.*

Table 2 Subcategorization frame examples

v	SCF	p(SCF|v)	% pass
acquire	ARG1#ARG2#	0.5461	0.1284
acquire	ARG1#ARG2#PP-in#	0.0886	0.0833
acquire	ARG1#ARG2#PP-from#	0.0406	0.1818
acquire	ARG1#ARG2#PP-by#	0.0406	0.0000
acquire	ARG1#ARG2#PP-during#	0.0295	0.3750

Table 3 Subcategorization slot examples

v	DEP	ll	% pass
acquire	ARG2#	579.96392	0.1512915
acquire	WH-when#	25.703417	0.1
acquire	PP-from#	22.716082	0.3333333
acquire	PP-by#	13.626654	0
acquire	PP-in#	13.416025	0.1666667

7. Comparison to existing lexicons

Several existing large-scale dictionaries and lexicons accommodate biological terms. Among them, many researchers use WordNet and the Specialist Lexicon for their text processing. WordNet is a general English resource which contains domain specific terms. The Specialist Lexicon was created by the National Library of Medicine, targeting the biomedical domain in general.

This section shows that our lexicon complements these popular lexical resources, by focusing on the words and relations that are

covered by our lexicon but not by these existing ones.

7.1 WordNet

WordNet (Fellbaum, 1998) is a general English thesaurus which additionally covers biological terms. We used WordNet 3.0[1] to evaluate term coverage.

Figure 2 shows the proportion of terminological words and relations (such as the word *retroregulate* and the relation *retroregulate* → *retroregulation*) in our lexicon that are also found in WordNet.

Since WordNet is not targeted at the biology domain, many biological terms and derivational relations are not listed.

7.2 UMLS Specialist Lexicon

The Specialist Lexicon[2] is a syntactic lexicon of biomedical and general English words, providing linguistic information about individual vocabulary items (Browne *et al*., 2003). Whilst it contains a large number of biomedical terms, our lexicon is tailored to the biology domain and covers more terms used within the biology domain, especially the molecular biology domain, than the Specialist Lexicon.

Figure 3 shows the proportion of words in our lexicon that are covered by the Specialist Lexicon.

Because the Specialist Lexicon is a biomedical lexicon and the target is broader than our lexicon, some biology-oriented words and relations are missing. For example the Specialist Lexicon includes the term *retro-regulator* but not *retro-regulate.* This means that derivational relations of *retro-regulate* are not covered by the Specialist Lexicon.

8. Conclusion and remarks

This paper has presented the BioLexicon, a unique resource comprising rich linguistic information suitable for bio-text mining applications. The lexicon has the following types of entries.

(1) Terminologies
(2) Derivational relations
(3) General English words
(4) Verb subcategorization frames

Comparisons with WordNet and the NLM Specialist Lexicon reveal that the BioLexicon covers words and relations which are pertinent to the biology domain but not included in these resources. We believe that it is a unique resource within the domain, which will play a complementary role to existing lexicons and thesauri.

The BioLexicon is available for non-commercial purposes under the Creative Commons license.

Our future work includes incorporating semantic event frames, such as gene regulation event frames, in the lexicon. Extrinsic evaluations of the lexicon in information extraction and question answering tasks are also planed.

Acknowledgement

This research is supported by EC IST project FP6-028099 (BOOTStrep), whose Manchester team is hosted by the JISC/BBSRC/EPSRC sponsored National Centre for Text Mining. The authors would like to thank Philip Cotter and Yoshimasa Tsuruoka for their assistance with the production of the lexical items. The authors also would like to thank Alessandro Lenci, Simone Marchi and Vito Pirrelli who contributed to the subcategorization extraction task.

References

Ananiadou, Sophia and John McNaught, editors. 2006. *Text Mining for Biology and Biomedicine*. Artech House, Norwood, MA.

Beisswanger, E., V. Lee, JJ Kim, D. Rebholz-Schuhmann, A. Splendiani, O. Dameron, S. Schulz, and U. Hahn. 2008. Gene Regulation Ontology (GRO): Design Principles and Use Cases. *Studies in health technology and informatics.* 136:9-14.

Browne, A.C., G. Divita, A.R. Aronson, and A.T. McCray. 2003. UMLS Language and Vocabulary Tools. In *Proc. of AMIA Annual Symposium 2003*, p.798.

Comrie, Bernard and Sandra A. Thompson. 2007. Lexical Normalization. In Timothy Shopen, editor, *Language Typology and Syntactic Description: Grammatical Categories and the Lexicon.* Chapter 8. Cambridge University Press.

[1] http://wordnet.princeton.edu/3.0/WordNet-3.0.tar.gz
[2] http://SPECIALIST.nlm.hih.gov

Dunning, T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61-74.

Fellbaum, C., editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA..

Francopoulo, G., M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. 2006. Lexical Markup Framework (LMF). In *Proc. of LREC 2006*, Genova, Italy.

Hara, Tadayoshi, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Adapting a Probabilistic Disambiguation Model of an HPSG Parser to a New Domain. In *Proc. of IJCNLP*, pages 199-210.

Kim, J-D., T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA Corpus - Semantically Annotated Corpus for Bio-Text Mining. *Bioinformatics*, 19:i180-i182.

Kim, J-D., T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA, In *Proc. of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 70-75.

Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data. In *Proc. of the Eighteenth International Conference on Machine Learning (ICML-2001),* pages 282-289.

Miyao, Yusuke and Jun'ichi Tsujii. 2003. Probabilistic modeling of argument structures including non-local dependencies. In *Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 285-291.

Pezik, P., A. Jimeno, V. Lee, and D. Rebholz-Schuhmann. 2008. Static Dictionary Features for Term Polysemy Identification. In *Proc of LREC-08 Workshop on Building and Evaluating Resources for Biomedical Text Mining*.

Quochi, Valeria, Monica Monachini, Riccardo Del Gratta, and Nicoletta Calzolari. 2008. A Lexicon for Biology and Bioinformatics: the BOOTStrep Experience. In *Proc. of Language Resources and Evaluation Conference (LREC-08)*, pages 28-30.

Santorini, Biatrice. 1990. *Part-of-Speech Tagging Guidelines for Penn Treebank Project.* 3rd Revision, 2nd Printing. ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz)

Sasaki, Yutaka, Yoshimasa Tsuruoka, John McNaught, and Sophia Ananiadou. 2008. How to Make the Most of NE Dictionaries in Statistical NER. *ACL-2008 Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP-08)*, pages 63-70.

Smith, L., T. Rindflesch, and W. J. Wilbur. 2004. MedPost: a Part-of-Speech Tagger for BioMedical Text. *Bioinformatics,* 20:2320-2321.

Tjong Kim Sang, Erik F. and J. Veenstra. 1999., Representing Text Chunks. In *Proc. of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99),* pages 173-179.

Tsai, R.T-H., W-C. Chou, Y-S. Su, Y-C. Lin, C-L. Sung, H-J. Dai, I. T-H. Yeh, W. Ku, T-Y. Sung, and W-L. Hsu. 2007. BIOSMILE. *BMC Bioinformatics*, 8:325.

Tsuruoka, Yoshimasa, John McNaught, Jun'ichi Tsujii, and Sophia Ananiadou. 2007. Learning String Similarity Measures for Gene/Protein Name Dictionary Look-up Using Logistic Regression. *Bioinformatics*, 23(20):2768-2774.

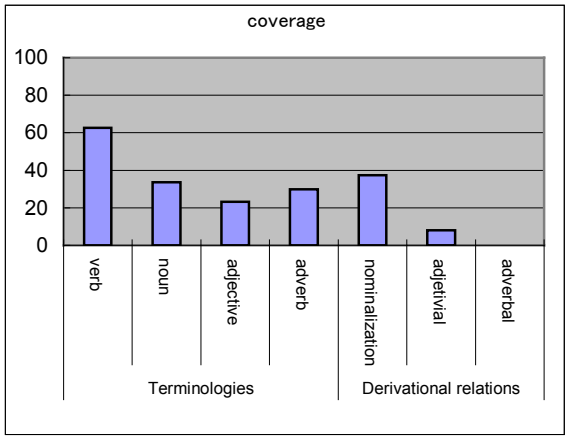

Figure 2 Word and relation coverage (%) in WordNet

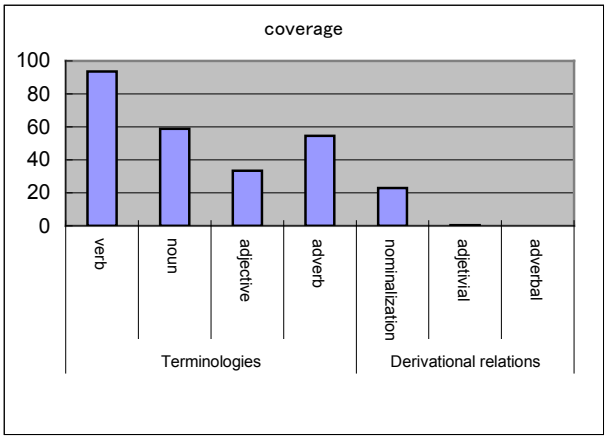

Figure 3 Word and Relation Coverage (%) in the Specialist Lexicon