Three BioNLP Tools Powered by a Biological Lexicon

Yutaka Sasaki [1] **Paul Thompson** [1] **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,Paul.Thompson,John.McNaught,Sophia.Ananiadou}@manchester.ac.uk

Abstract

In this paper, we demonstrate three NLP applications of the BioLexicon, which is a lexical resource tailored to the biology domain. The applications consist of a dictionary-based POS tagger, a syntactic parser, and query processing for biomedical information retrieval. Biological terminology is a major barrier to the accurate processing of literature within biology domain. In order to address this problem, we have constructed the BioLexicon using both manual and semi-automatic methods. We demonstrate the utility of the biology-oriented lexicon within three separate NLP applications.

1 Introduction

Processing of biomedical text can frequently be problematic, due to the huge number of technical terms and idiosyncratic usages of those terms. Sometimes, general English words are used in different ways or with different meanings in biology literature.

There are a number of linguistic resources that can be use to improve the quality of biological text processing. WordNet (Fellbaum, 1998) and the NLP Specialist Lexicon [1] are dictionaries commonly used within biomedical NLP.

WordNet is a general English thesaurus which additionally covers biological terms. However, since WordNet is not targeted at the biology domain, many biological terms and derivational relations are missing.

The Specialist Lexicon 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, its focus is on medical terms. Therefore some biology-specific terms, *e.g.,* molecular biology terms, are not the main target of the lexicon.

In response to this, we have constructed the BioLexicon (Sasaki *et al*., 2008), a lexical resource tailored to the biology domain. We will demonstrate three applications of the BioLexicon, in order to illustrate the utility of the lexicon within the biomedical NLP field.

The three applications are:

- BLTagger: a dictionary-based POS tagger based on the BioLexicon
- Enju full parser enriched by the BioLexicon
- Lexicon-based query processing for information retrieval

2. Summary of the BioLexicon

In this section, we provide a summary of the BioLexicon (Sasaki *et al*., 2008). It contains words belonging to four part-of-speech categories: verb, noun, adjective, and adverb.

Quochi *et al*.(2008) designed the database model of the BioLexicon which follows the Lexical Markup Framework (Francopoulo *et al*., 2008).

2.1 Entries in the Biology Lexicon

The BioLexicon accommodates both general English words and terminologies. Biomedical terms were gathered from existing biomedical databases. Detailed information regarding the sources of biomedical terms can be found in (Rebholz-Schuhmann *et al*., 2008). The lexicon entries consist of the following:

(1) Terminological verbs: 759 base forms (4,556 inflections) of terminological verbs with automatically extracted verb subcategorization frames

[1] http://SPECIALIST.nlm.hih.gov

Proceedings of the EACL 2009 Demonstrations Session, pages 61–64,
Athens, Greece, 3 April 2009.

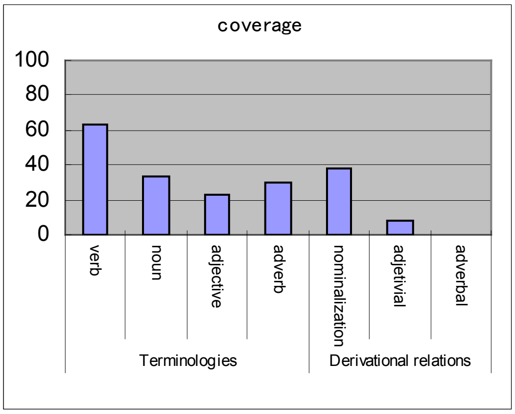

Fig. 1 Comparison with WordNet

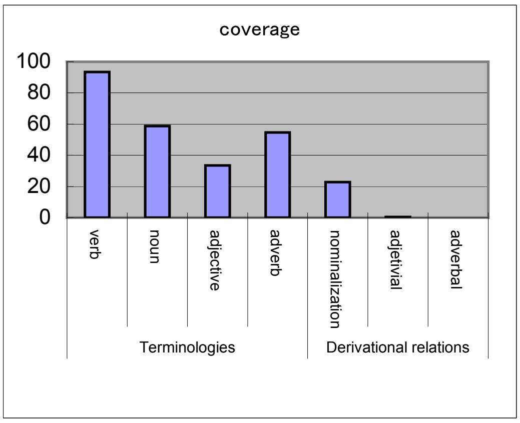

Fig. 2 Comparison with Specialist Lexicon

(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. (Sasaki *et al*., 2008)

2.2. Comparison to existing lexicons

This section focuses on the words and derivational relations of words that are covered by our BioLexicon but not by comparable existing resources. Figures 1 and 2 show the percentage of the terminological words and derivational relations (such as the word *retroregulate* and the derivational relation *retroregulate* → *retroregulation*) in our lexicon that are also found in WorNet and the Specialist Lexicion.

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

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.

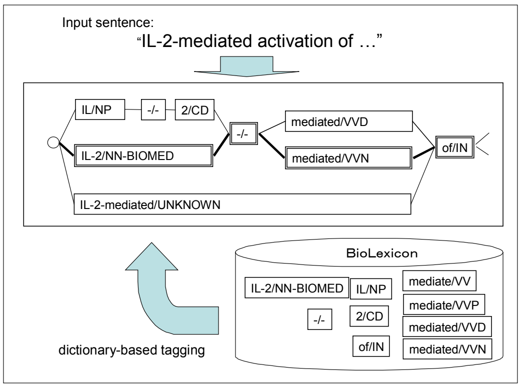

Fig. 3 BLTagger example

3. Application 1: BLTagger

Dictionary-based POS tagging is advantageous when a sentence contains technical terms that conflict with general English words. If the POS tags are decided without considering possible occurrences of biomedical terms, then POS errors could arise.

For example, in the protein name "met proto-oncogene precursor", *met* might be incorrectly recognized as a verb by a non dictionary-based tagger.

In the dictionary, biomedical terms are given POS tag "NN-BIOMED". Given a sentence, the dictionary-based POS tagger works as follows.

- Find all word sequences that match the lexical entries, and create a token graph (*i.e.,* trellis) according to the word order.
- Estimate the score of every path using the weights of the nodes and edges, through training using Conditional Random Fields.
- Select the best path.

Figure 3 shows an example of our dictionary-based POS tagger BLTagger.

Suppose that the input is “IL-2-mediated activation of”. A trellis is created based on the lexical entries in the dictionary. The selection criteria for the best path are determined by the CRF tagging model trained on the Genia corpus (Kim *et al*., 2003). In this example,

```
IL-2/NN-BIOMED -/- mediated/VVN
activation/NN of/IN
```

is selected as the best path.

Following Kudo et al. (2004), we adapted the core engine of the CRF-based morphological analyzer, MeCab[2], to our POS tagging task.

The features used were:

- *POS*
- BIOMED
- *POS*-BIOMED
- bigram of adjacent *POS*
- bigram of adjacent BIOMED
- bigram of adjacent *POS*-BIOMED

During the construction of the trellis, white space is considered as the delimiter unless otherwise stated within dictionary entries. This means that unknown tokens are character sequences without spaces.

As the BioLexicon associates biomedical semantic IDs with terms, the BLTagger attaches semantic IDs to the tokenizing/tagging results.

4. Application 2: Enju full parser with the BioLexicon

Enju (Miyao, *et al*., 2003) is an HPSG parser, which is tuned to the biomedical domain. Sentences are parsed based on the output of the Stepp POS tagger, which is also tuned to the biomedical domain.

To further tune Enju to the biology domain, (especially molecular biology), we have modified Enju to parse sentences based on the output of the BLTagger.

As the BioLexicon contains many multi-word biological terms, the modified version of Enju parses token sequences in which some of the tokens are multi-word expressions. This is effective when very long technical terms (*e.g.*, more than 20 words) are present in a sentence.

To use the dictionary-based tagging for parsing, unknown words should be avoided as much as possible. In order to address this issue, we added entries in WordNet and the Specialist Lexicion to the dictionary of BLTagger.

The enhancement in the performance of Enju based on these changes is still under evaluation. However, we demonstrate a functional, modified version of Enju.

5. Application 3: Query processing for IR

It is sometimes the case that queries for biomedical IR systems contain long technical terms that should be handled as single multi-word expressions.

We have applied BLTagger to the TREC 2007 Genomics Track data (Hersh *et al*., 2007). The goal of the TREC Genomics Track 2007 was to generate a ranked list of passages for 36 queries that relate to biological events and processes.

Firstly, we processed the documents with a conventional tokenizer and standard stop-word remover, and then created an index containing the words in the documents. Queries are processed with the BLTagger and multi-word expressions are used as phrase queries. Passages are ranked with Okapi BM25 (Robertson *et al.,* 1995).

Table 1 shows the preliminary Mean Average Precision (MAP) scores of applying the BLTagger to the TREC data set.

By adding biology multi-word expressions identified by the BLTagger to query terms (row (a)), we were able to obtain a slightly better Passage2 score. As the BLTagger outputs semantic IDs which are defined in the BioLexicon, we tried to use these semantic IDs for query expansion (rows (b) and (d)). However, the MAP scores degraded.

[2] http://sourceforge.net/project/showfiles.php?group id=177856/

Table 1 Preliminary MAP scores for TREC Genomics Track 2007 data

Query expansion method	Passage2 MAP	Aspect MAP	Document MAP
(a) BioLexicon terms	0.0702	0.1726	0.2158
(b) BioLexicon terms + semantic IDs	0.0696	0.1673	0.2148
(c) no query expansion (baseline)	0.0683	0.1726	0.2183
(d) semantic IDs	0.0677	0.1670	0.2177

6. Conclusions

We have demonstrated three applications of the BioLexicon, which is a resource comprising linguistic information, targeted for use within bio-text mining applications.

We have described the following three applications that will be useful for processing of biological literature.

- BLTagger: dictionary-based POS tagger based on the BioLexicon
- Enju full parser enriched by the BioLexicon
- Lexicon-based query processing for information retrieval

Our future work will include further intrinsic and extrinsic evaluations of the BioLexicon in NLP, including its application to information extraction tasks in the biology domain. The BioLexicon is available for non-commercial purposes under the Creative Commons license.

Acknowledgements

This research has been supported by the EC IST project FP6-028099 (BOOTStrep), whose Manchester team is hosted by the JISC/BBSRC/EPSRC sponsored National Centre for Text Mining.

References

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.

Dietrich Rebholz-Schuhmann, Piotr Pezik, Vivian Lee, Jung-Jae Kim, Riccardo del Gratta, Yutaka Sasaki, Jock McNaught, Simonetta Montemagni, Monica Monachini, Nicoletta Calzolari, Sophia Ananiadou, BioLexicon: Towards a Reference Terminological Resource in the Biomedical Domain, *the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB-2008)* (Poster), Toronto, Canada, 2008. (http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/BioLexicon_Poster_EBI_UoM_ILC.pdf)

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.

Hersh, W., Aaron Cohen, Lynn Ruslen, and Phoebe Roberts, TREC 2007 Genomics Track Overview, *TREC-2007*, 2007.

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

Kudo T., Yamamoto K., Matsumoto Y., Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP-04)*, pp. 230–237, 2004.

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, Y. and J. 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.

Quochi, V., Monachini, M., Del Gratta, R., Calzolari, N., A lexicon for biology and bioinformatics: the BOOTStrep experience. In *Proc. of LREC 2008*, Marrakech, 2008.

Robertson, S.E., Walker S., Jones, S., Hancock-Beaulieu M.M., and Gatford, M., 1995. Okapi at TREC-3. In *Proc of Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126.

Yutaka Sasaki, Simonetta Montemagni, Piotr Pezik, Dietrich Rebholz-Schuhmann, John McNaught, and Sophia Ananiadou, BioLexicon: A Lexical Resource for the Biology Domain, In *Proc. of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, 2008.