Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, July 2002, pp. 9-13. Association for Computational Linguistics.

Tagging Gene and Protein Names in Full Text Articles

Lorraine Tanabe and W. John Wilbur
National Center for Biotechnology Information
NLM, NIH
Bethesda, Maryland 20894

Abstract

Current information extraction efforts in the biomedical domain tend to focus on finding entities and facts in structured databases or MEDLINE® abstracts. We apply a gene and protein name tagger trained on Medline abstracts (ABGene) to a randomly selected set of full text journal articles in the biomedical domain. We show the effect of adaptations made in response to the greater heterogeneity of full text.

1 Introduction

The application of large-scale genomics and proteomics technologies towards a wide variety of biological questions has resulted in a continuous stream of information regarding thousands of genes and gene products into the Medline database of biomedical abstracts. This repository has been recognized as a rich knowledge source for biological information retrieval, information extraction and text mining. However, abbreviated scientific abstracts cannot contain the same volume of information as the full text articles that they represent. It was recently shown that only 30% of protein interactions contained in the Dictionary of Interacting Proteins (DIP) (Xenarios et al., 2000) could be found in Medline sentences containing DIP protein pairs (Blaschke et al., 2000). This finding suggests that current information extraction efforts being applied to biomedical abstracts should be extended to full text databases.

The basic task of identifying gene and protein names is a necessary first step towards making full use of the information encoded in biomedical text. This remains a challenging task due to the irregularities and ambiguities in gene and protein nomenclature. The irregularities are mainly the result of a lack of naming conventions, as well as the widespread practice of using many synonyms for one gene or protein. A glance at the *Nomenclature section* of the *Nature Genetics* website (http://www.nature.com/ng/web_specials/nomen/) shows the scope of the problem, as well as ideas for addressing it. The nomenclature guidelines implore authors to consult relevant nomenclature committees before announcing new genes, and to provide synonyms for genes in abstracts. Additional rules specify that:

4. Gene symbols are always italicised and never contain hyphens, greek letters, roman numerals, subscripts or superscripts.
5. All letters in human genes are upper-case…all letters in mouse genes are lower-case…

Unfortunately, we are currently at a stage where these types of rules are not consistently applied to most biomedical abstracts, let alone to full text documents. Until the biomedical community adheres uniformly to nomenclature guidelines, ambiguities regarding gene/protein names will continue to be an obstacle for natural language processing of biomedical text. These ambiguities become apparent at the morphological, syntactic and semantic levels. For example, *caco*-2 refers to a cell line, but *pai*-1 is a gene name. Gene and protein names can contain verbs and other parts of speech that are hard to distinguish from the surrounding text, as in *deleted in azoospermia-like*, *son of sevenless, ran, man, young arrest* and *never in mitosis*. Genes can be transfected into cells, or combined with chemicals, resulting in ambiguous terms like *CHO-A(3)* and *ca^{2+}/calmodulin*. The semantic notion of a gene or protein is quite arbitrary – is *ACTTGGAATGACC* a gene name? In addition to sequences, there are mutations, motifs, receptors, antibodies, hormones, channels,

chromosomal locations and disease loci to consider. The domain-specific irregularities and ambiguities just described are superimposed upon the ambiguities in the natural language itself, so it is not surprising that the identification of gene and protein names in biomedical text remains a difficult and challenging task. The methodologies applied to this fundamental problem include rule-based and/or pattern matching methods (Fukuda et al., 1998) (Thomas et al., 2000) (Yoshida et al., 2000) (Jenssen et al., 2001) (Ono et al., 2001) (Yu at al, 2002) (Bunescu et al., 2002), a modified BLAST algorithm (Krauthammer et al., 2000), Hidden Markov Models (HMMs) (Collier et al., 2000) (Proux et al., 1998), Naive Bayes and decision trees (Nobata et al., 1999), under specified parsing with knowledge sources (Rindflesch et al, 2000), and context-free grammars (Gaizauskas, 2000).

In this paper, we evaluate the application of a gene and protein name tagger trained on Medline abstracts (ABGene) (Tanabe and Wilbur, 2002) to a randomly selected set of 1,000 PUBMEDCENTRAL™ (PMC) articles. PMC is a digital archive of full text peer-reviewed biomedical articles launched in February 2000 by the National Center for Biotechnology Information (NCBI) and the U.S. National Library of Medicine (NLM®) (Roberts et al., 2001). We present two adaptations made in response to the greater heterogeneity of full text, and evaluate how they affect the performance of ABGene on a test set of 2600 full text sentences.

2 Methods

We first give an overview of ABGene's method for extracting gene and protein names from biomedical citations, and then present some modifications to ABGene designed to improve its performance on full text articles.

2.1 ABGene Overview

We previously trained the Brill POS tagger (Brill, 1994) to recognize protein and gene names in biomedical text using a training set of 7,000 Medline sentences. We updated the lexicon included in the Brill package (Brown Corpus plus Wall Street Journal corpus) with entries from the UMLS® SPECIALIST lexicon (McCray et al. 1994, Humphreys et al. 1998), and generated a list of bigrams and a word list from all of MEDLINE to customize the training for our purposes. ABGene processing begins by using these automatically generated rules from the Brill tagger to extract single word gene and protein names from biomedical abstracts (see Table 1).

Lexical Rule	**Description**
NNP gene fgoodleft GENE	*Change the tag of a word from NNP to GENE if the word gene can appear to the right*
-A hassuf 2 GENE	*Change the tag of a word from anything to GENE if it contains the suffix -A*
c- haspref 2 GENE	*Change the tag of a word from anything to GENE if it contains the prefix c-*
GENE cell fgoodright NNP	*Change the tag of a word from GENE to NNP if the word cell can appear to the left*
Contextual Rule	**Description**
NNP GENE PREV1OR2WD genes	*Change the tag of a word from NNP to GENE if one of the two preceding words is genes*
NNP GENE NEXTBIGRAM (GENE	*Change the tag of a word from NNP to GENE if the two following words are tagged (and GENE*
CD GENE SURROUNDTAG CC)	*Change the tag of a word from CD to GENE if the preceding word is tagged CC and the following word is tagged)*
VBG JJ NEXTTAG GENE	*Change the tag of a word from VBG to JJ if the next word is tagged GENE*

Table 1. Examples of lexical and contextual rules learned by the Brill tagger. NNP = proper noun, CD = cardinal number, CC = coordinating conjunction, JJ = adjective, VBG = verb, gerund/present participle

This is followed by extensive filtering for false positives and false negatives. A key step during the filtering stage is the extraction of multi-word gene and protein names that are prevalent in the literature but inaccessible to the Brill tagger.

During the false positive filtering step, the GENE tag is removed from a word if it matches a term from a list of 1,505 precompiled general biological terms (acids, antagonist, assembly, antigen, etc.), 39 amino acid names, 233 restriction enzymes, 593 cell lines, 63,698 organism names from the NCBI Taxonomy

Database (Wheeler et al. 2000) or 4,357 non-biological terms. Non-biological terms were obtained by comparing word frequencies in MEDLINE versus the Wall Street Journal (WSJ) using the following expression, where p is the probability of occurrence:

log(p(word occurs in MEDLINE)/ p(word occurs in WSJ))< 1

Additional false positives are found by regular expressions including numbers followed by measurements (25 mg/ml) and common drug suffixes (-ole, -ane, -ate, -ide, -ine, -ite, -ol, -ose, cooh).

The false negative filter recovers a single word name if it: 1) matches a list of 34,555 single word names and 7611 compound word names compiled from LocusLink (Pruitt & Maglott 2001) and the Gene Ontology Consortium (2000) (Wain et al., 2002) and contains a good context word before or after the name, or 2) contains a low frequency trigram and a good context word before or after the name. The context words were automatically generated by a probabilistic algorithm, using the LocusLink/Gene Ontology set and a large collection of texts in which these gene names occur. We computed a log odds score or Bayesian weight for all non-gene name words indicating their propensity to predict an adjacent gene name in the texts.

Compound word names are recovered using terms that occur frequently in known gene names. Recombination of these terms produce compound words that also tend to be gene/protein names. These terms include the digits 1-9, the letters a-z, the roman numerals, the Greek letters, functional descriptors (*adhesion)*, organism identifiers (*hamster)*, activity descriptors *(promoting)*, placement indicators *(early)*, and generic descriptors *(light)*. In addition to the 415 exact terms, we added regular expressions that allow for partial matches or special patterns such as words without vowels, words with numbers and letters, words in capital letters, and common prefixes and suffixes *(-gene, -like, -ase)*.

Finally, Bayesian learning (Langley 1996, Mitchell 1997, Wilbur 2000) is applied to rank documents by similarity to documents with known gene/protein names. Documents below a certain threshold are considered to have no gene/protein names in them.

2.2 Modifications for Full Text Articles

The full text PMC articles are longer than abstracts, and contain extraneous information like grant numbers and laboratory reagents, along with figures and tables. An attempt to take windows of varying sizes of the full text in order to rank the windows by similarity to abstracts with known gene names was unsuccessful. High scoring windows often hid false positives, and low scoring windows could contain gene and protein name contexts infrequently encountered in Medline abstracts. However, we determined that the classifier could be used on the sentence level for full text articles, and show the effect of an assumption that sentences below a zero threshold do not contain gene/protein names.

We tried to increase the performance of ABGene on the PMC articles by adding a final processing step. We ran ABGene on 2.16 million Medline abstracts similar to documents with known gene names, and extracted 2.42 million unique gene/protein names. We counted the number of times each unique name was given the GENE tag by ABGene in the 2.16 million abstracts, and then extracted three groups of putative gene/protein names from this large set, with count thresholds at 10 (134,809 names), 100 (13,865 names) and 1000 (1136 names).

During the final stage of processing, terms in sentences with scores greater than 2 are checked against these lists of supposed gene/protein names. We show the effect of tagging terms with counts of at least 10, 100 and 1000 in the putative gene/protein list.

3 Experiment and Results

We evaluated the performance of ABGene on 2600 PMC sentences from 13 score levels ranging from –8 to 60+. No attempt was made to narrow the set using query terms. The sentences were selected as follows: half of the test set consists of the first 100 sentences from each score level, and the other half consists of 100 sentences selected at random from each score level. Precision and recall results are shown for each individual score range in Table 2, and cumulative results are shown in Table 3. The number of words tested varies for each score level because longer sentences tend to have higher scores. Also, sentences with scores near zero tend to be table or figure entries, with only a few words each.

Score Range	#words tested	TP + FN	P	R	P 1000	R 1000	P 100	R 100	P 10	R 10
60+	13,442	1347	0.742	0.640	0.726	0.667	0.686	0.692	0.603	0.716
30 to 60	7,953	530	0.672	0.638	0.673	0.667	0.649	0.699	0.590	0.765
20 to 30	6,392	401	0.757	0.646	0.751	0.671	0.708	0.748	0.624	0.801
15 to 20	5,508	302	0.722	0.593	0.719	0.619	0.672	0.659	0.561	0.735
10 to 15	5,100	269	0.755	0.688	0.743	0.710	0.681	0.747	0.579	0.792
8 to 10	4,618	226	0.707	0.588	0.689	0.637	0.615	0.686	0.512	0.770
6 to 8	4,327	170	0.703	0.571	0.692	0.594	0.641	0.641	0.479	0.724
4 to 6	4,054	122	0.571	0.590	0.562	0.631	0.500	0.648	0.392	0.713
2 to 4	3,667	59	0.541	0.559	0.508	0.559	0.404	0.610	0.270	0.644
0 to 2	1,551	9	0.200	0.444	0.200	0.444	0.200	0.444	0.200	0.444
-2 to 0	4,595	0	no tp	no tp	no tp	no tp	no tp	no tp	no tp	no tp
-4 to -2	5,299	1	0.040	1.000	0.040	1.000	0.040	1.000	0.040	1.000
-8 to -4	5,495	0	no tp	no tp	no tp	no tp	no tp	no tp	no tp	no tp

Table 2. Precision and recall for each score range. TP+FN = number of gene names; P = precision without final step, R = recall without final step, P 1000 = precision with 1000 count threshold at final step, R 1000 = recall with 1000 count threshold at final step, P 100 = precision with 100 count threshold at final step, R 100 = recall with 100 count threshold at final step, P 10 = precision with 10 count threshold at final step, R 10 = recall with 10 count threshold at final step.

SCORE	P	R	P 1000	R 1000	P 100	R 100	P 10	R 10
60	0.742	0.251	0.726	0.261	0.686	0.273	0.603	0.283
30	0.721	0.349	0.710	0.364	0.675	0.381	0.599	0.402
20	0.727	0.424	0.717	0.443	0.681	0.468	0.604	0.496
15	0.727	0.476	0.717	0.497	0.680	0.526	0.598	0.560
10	0.729	0.530	0.720	0.553	0.680	0.584	0.596	0.622
8	0.728	0.569	0.718	0.595	0.675	0.629	0.589	0.673
6	0.727	0.597	0.716	0.624	0.673	0.661	0.582	0.709
4	0.720	0.618	0.710	0.646	0.665	0.684	0.573	0.734
2	0.716	0.628	0.706	0.656	0.659	0.695	0.563	0.745
0	0.713	0.629	0.702	0.657	0.656	0.696	0.562	0.746

Table 3. Cumulative precision and recall using the score as a lower threshold.

3.1 Problematic Areas in Full Text

The false positive gene/protein names found in the PMC articles reveal new difficulties for the basic task of identifying gene and protein names in biomedical text. For example, in abstracts, entities like restriction enzyme sites, laboratory protocol kits, primers, vectors, molecular biology supply companies and chemical reagents are usually scarce. However, in the methods section of a full document, they appear regularly, adding to the morphological, syntactic and semantic ambiguities previously mentioned. Illustrative examples include *bio-rad, centricon-30 spin, xbai sites, mg2, geneamp* and *pgem3z*. A significant source of false negatives consists of tables and figures from full text, which completely lack contextual cues and/or indicator words. These problems can be addressed by eliminating processing of materials and methods sections, tables and figures. Another significant source of false negatives is an artifact of the PMC format, for example, *beta* is translated to *[beta]*, thus a name like *beta1 integrin* becomes *[beta]1 integrin* in PMC. This is easily addressed by removing the PMC formatting prior to processing, and has already been completed for future work on PMC articles.

4 Conclusion

We conclude that an information extraction system to tag gene and protein names in Medline abstracts (ABGene) can be applied to full text articles in the biomedical domain. We

have shown how modifications to the processing (applying a sentence score threshold, and using a large pool of putative gene/protein names) can affect the system's performance. We are currently exploring methods to filter the 2.16 million putative gene/protein names extracted from Medline using our system. The resulting set of gene/protein names, a significant addition to the 42K names available from the Gene Ontology Consortium and LocusLink, will be used to improve the performance of text processing on full text articles in the biomedical domain.

References

Blaschke, C. and Valencia, A. (2001) Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. *Comparative and Functional Genomics,* **2**, 196-206.

Brill, Eric. (1994) Some advances in transformation-based part of speech tagging. In *Proceedings of the National Conference on Artificial Intelligence.* AAAI Press, pp. 722-727.

Bunescu, R., Ge, R., Mooney, R.J., Marcotte, E., and Ramani, A.K. (2002) Extracting gene and protein names from biomedical abstracts. http://www.cs.utexas.edu/users/ml/publication/ie.html.

Collier, N., Nobata, C., and Tsujii, J. (2000) Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, pp. 201-207.

Fukuda, K., Tsunoda. T., Tamura, A. and Takagi. T. (1998) Toward information extraction: identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing (PSB98),* pp. 705-716.

The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.,* **25**, 25-29.

Humphreys K., Demetriou G., and Gaizauskas, R. (2000) Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. In *Proceedings of the Pacific Symposium on Biocomputing (PSB2000)* , pp. 502-513.

Jenssen, T., Laegreid, A., Kormorowski, J., and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet.,* **28**, 21-28.

Krauthammer, M., Rzhetsky, A., Morozov, P., and Friedman, C. (2000) Using BLAST for identifying gene and protein names in journal articles. *Gene*, **259**, 245-252.

Langley, P. (1996) *Elements of Machine Learning.* Morgan Kaufmann Publishers, Inc., San Francisco.

McCray, A.T., Srinivasan, S. and Browne, A. C. Lexical methods for managing variation in biomedical terminologies. In *SCAMC* '94, pp. 235-239.

Mitchell, T. M. (1997) *Machine Learning.* WCB/McGraw-Hill, Boston.

Nobata, C., Collier, N., and Tsujii, J. (1999) Automatic term identification and classification in biology texts**.** In *Proceedings of the Natural Language Pacific Rim* Symposium, pp. 369-374.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001) Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, **17**, 155-161.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V., and Jacq, B. (1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. In *Proceedings of the Ninth Workshop on Genome Informatics*, pp. 72-80.

Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.,* **29**, 137-140.

Rindflesch, T. C., Tanabe, L., Weinstein, J. W., and Hunter, L. (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of the Pacific Symposium on Biocomputing (PSB2000)*, pp. 514-525.

Roberts, R.J., Varmus, H.E., and Ashburner, M. (2001) Information access: building a Genbank of the published literature. *Science*, **291**, 2318-2319.

Tanabe, L., and Wilbur, W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, in press.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroh, M. (2000) Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB2000)*, pp. 541-552.

Wain, H. M., Lush, M., Ducluzeau, F. , and Povey, S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.,* **30**, 169-171.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., and Rapp, B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28***,* 10-14.

Wilbur, W. J. (2000) Boosting naive bayesian learning on a large subset of MEDLINE. In *American Medical Informatics 2000 Annual Symposium*, Los Angeles, CA, pp. 918-922.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., and Eisenberg, D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.,* **28**, 289-291.

Yoshida, M., Fukuda, K., and Takagi, T. (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, **16**, 169-175.

Yu, H., Hripcsak, G., and Friedman, C. (2002) Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc*., **9**, 262-272.