MedTag: A Collection of Biomedical Annotations

L.H. Smith[†], L. Tanabe[†], T. Rindflesch[‡], W.J. Wilbur
National Center for Biotechnology Information
[‡]Lister Hill National Center for Biomedical Communications
NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894
{lsmith,tanabe,wilbur}@ncbi.nlm.nih.gov
rindflesch@nlm.nih.gov

Abstract

We present a database of tagged biomedical text corpora merged into a portable data structure with uniform conventions. The data are available in flat files along with software to facilitate loading the data into a relational SQL database.[1]

1 Introduction

Annotated text corpora are used in modern computational linguistics research and development to fine-tune computer algorithms for analyzing and classifying texts and textual components. Two important factors for useful text corpora are 1) accuracy and consistency of the annotations, and 2) usability of the data. We have recently updated the text corpora we use in our research with respect to these criteria.

Three different corpora were combined. The ABGene corpus consists of over 4 000 sentences annotated with gene and protein named entities. It was originally used to train the ABGene tagger to recognize gene/protein names in MEDLINE records, and recall and precision rates in the lower 70 percentile range were achieved (Tanabe and Wilbur, 2002). The MedPost corpus consists of 6 700 sentences, and is annotated with parts of speech, and gerund arguments. The MedPost tagger was trained on 3 700 of these sentences and achieved an accuracy of 97.4% on the remaining sentences (Smith et. al., 2004). The GENETAG corpus for gene/protein named entity identification, consists of 20 000 sentences and was used in the BioCreative 2004 Workshop (Yeh et. al., 2005; Tanabe et. al., 2005) (only 15 000 sentences are currently released, the remaining 5 000 are being retained for possible use in a future workshop). Training on a portion of the data, the top performing systems achieved recall and precision rates in the lower 80 percentile range. Because of the scarcity of good annotated data in the realm of biomedicine, and because good performance has been obtained using this data, we feel there is utility in presenting it to a wider audience.

All of the MedTag corpora are based on MEDLINE abstracts. However, they were queried at different times, and used different (but similar) algorithms to perform tokenization and sentence segmentation. The original annotations were assigned to tokens, or sequences of tokens, and extensively reviewed by the authors at different times for the different research projects.

The main goals in combining and updating these corpora into a single corpus were to

1. update the text for all corpora to that currently found in MEDLINE, storing a correct citation and the original, untokenized text for each excerpt
2. eliminate tokenization dependence
3. put all text and annotations into a common database format
4. provide programs to convert from the new corpus format to the data formats used in previous research

[1] `ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedTag/medtag.tar.gz`

2 Updating the Corpora

We describe what was done to update the original corpora, locating original sources and modifying the text where needed.

2.1 Source Data

The original data of the three corpora were assembled and the text was used to search MEDLINE to find the closest match. An exact or near exact match was found for all but a few excerpts. For only a few excerpts, the MEDLINE record from which the excerpt was originally taken had been removed or modified and an alternative sentence was selected. Thus, each excerpt in the database is taken from a MEDLINE record as it existed at one time in 2004. In order to preserve the reference for future work, the PubMed ID and citation data were also retrieved and stored with each excerpt. Each excerpt in the current database roughly corresponds to a sentence, although the procedure that extracted the sentence is not specified.

2.2 Tokenization Dependence

In the original ABGene and GENETAG corpora, the gene and protein phrases were specified by the tokens contained in the phrase, and this introduced a dependence on the tokenization algorithm. This created problems for researchers who wished to use a different tokenization. To overcome this dependence, we developed an alternative way of specifying phrases. Given the original text of an excerpt, the number of non-whitespace characters to the start of the phrase does not depend on the tokenization. Therefore, all annotations now refer to the first and last character of the phrase that is annotated. For example the protein *serum LH* in the excerpt

> There was no correlation between *serum LH* and chronological or bone age in this age group, which suggests that the correlation found is not due to age-related parallel phenomena.

is specified as characters 28 to 34 (the first character is 0).

2.3 Data Model

There are two main record types in the database, EXCERPT and ANNOTATION. Each EXCERPT record stores an identifier and the original corpus code (abgene, medpost, and genetag) as well as subcorpus codes that were defined in the original corpora. The original text, as it was obtained from MEDLINE, is also stored, and a human readable citation to the article containing the reference.

Each ANNOTATION record contains a reference to the excerpt (by identifier and corpus), the character offset of the first and last characters of the phrase being annotated (only non-whitespace characters are counted, starting with 0), and the corresponding annotation. The annotated text is stored for convenience, though it can be obtained from the corresponding excerpt record by counting non-whitespace characters.

The data is provided as an ASCII file in a standard format that can be read and loaded into a relational database. Each record in the file begins with a line of the form >>*table_name* where *table_name* is the name of the table for that record. Following the table name is a series of lines with the form *field: value* where *field* is the name of the field and *value* is the value stored in that field.

Scripts are provided for loading the data into a relational database, such as *mysql* or *ORACLE*. SQL queries can then be applied to retrieve excerpts and annotations satisfying any desired condition. For example, here is an SQL query to retrieve excerpts from the MedPost corpus containing the token *p53* and *signaling* or *signalling*

```
select text from excerpt
where text like '%p53%'
and text rlike 'signa[l]*ing';
```

2.4 Programs

A web-based corpus editor was used to enter and review annotations. The code is being made available, as is, and requires that the data are loaded into a mysql database that can be accessed by a web server. The interface supports two annotation types: MedPost tags and arbitrary phrase annotations. MedPost tags are selectable from a pull-down menu of preprogrammed likely tags. For entering phrase annotations, the user highlights the desired phrase, and pressing the enter key computes and saves the first and last character offsets. The user can then enter the annotation code and an optional comment before saving it in the database. A screen dump of the

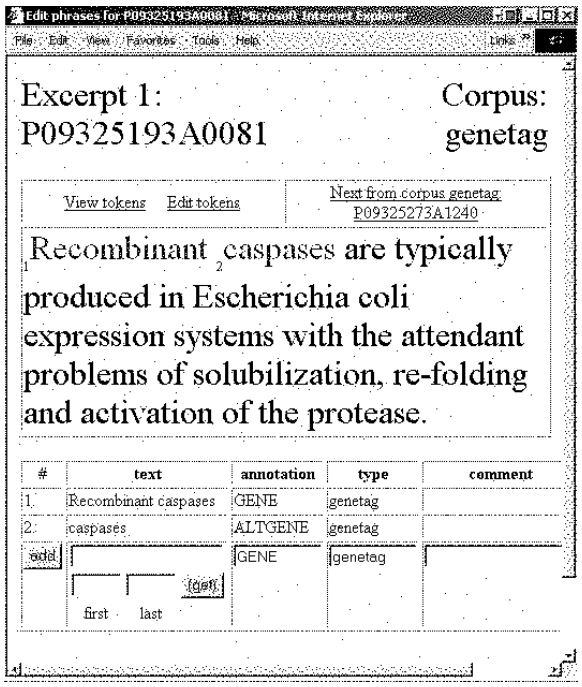

Figure 1: A screen capture of the annotator's interface and the GENETAG-05 annotations for a sentence.

phrase annotations for a sentence in the genetag corpus is shown in figure 1.

The data from the database was dumped to the flat file format for this release. We have also included some files to accommodate previous users of the corpora. A perl program, `alt_eval.perl` is included that replaces the GENETAG evaluation program using non-whitespace character numbers instead of token numbers. Copies of the ABGene and MedPost corpora, in the original formats, are also included.

3 GENETAG-05

We are including a new version of GENETAG, GENETAG-05, as part of the MedTag system. GENETAG-05 differs from GENETAG in four ways: 1) the definition of a gene/protein entity has been modified, 2) significant annotation errors in GENETAG have been corrected, 3) the concept of a non-specific entity has been refined, and 4) character-based indices have been introduced to reduce tokenization problems. We believe that these changes result in a more accurate and robust corpus.

GENETAG-05 maintains a wide definition of a gene/protein entity including genes, proteins, domains, sites, sequences, and elements, but excluding plasmids and vectors. The specificity constraint requires that a gene/protein name must be included in the tagged entity. This constraint has been applied more consistently in GENETAG-05. Additionally, plain sequences like *ATTGGCCTTTAAC* are no longer tagged, embedded names are tagged (*ras*-mediated), and significantly more terms have been judged to violate the specificity constraint (*growth factor, proteases, protein kinase, ribonuclease, snoRNA, rRNA, tissue factor, tumor antigen, complement, hormone receptors, nuclear factors, etc.*).

The original GENETAG corpus contains some entities that were erroneously tagged as gene/proteins. Many of these errors have been corrected in the updated corpus. Examples include *camp-responsive elements, mu element, VDRE, melanin, dentin, myelin, auxin, BARBIE box, carotenoids,* and *cellulose.* Error analysis resulted in the updated annotation conventions given in Table 1.

Enzymes are a special class of proteins that catalyze biochemical reactions. Enzyme names have varying degrees of specificity, so the line drawn for tagging purposes is based on online resources[2] as well as background knowledge. In general, tagged enzymes refer to more specific entities than untagged enzymes (*tyrosine kinase* vs. *protein kinase*, *ATPase* vs. *protease*). Enzymes that can refer to either DNA or RNA are tagged if the reference is specified (*DNA endonuclease* vs. *endonuclease*). Enzymes that do not require DNA/RNA distinction are tagged (*lipase* vs. *ligase*, *cyclooxygenase* vs. *methylase*). Non-specific enzymes are tagged if they clearly refer to a gene or protein, as in (1).

> 1) The structural gene for *hydrogenase* encodes a protein product of molecular mass 45820 Da.

Semantic constraints in GENETAG-05 are the same as those for GENETAG. To illustrate, the name in (2) requires *rabies* because *RIG* implies that the gene mentioned in this sentence refers to the *rabies*

[2] http://cancerweb.ncl.ac.uk/omd/copyleft.html
http://www.onelook.com/

immunoglobulin, and not just any *immunoglobulin*. In (3), the word *receptor* is necessary to differentiate *IGG receptor* from *IGG*, a crucial biological distinction. In (4), the number *1* is needed to accurately describe a specific type of *tumor necrosis factor*, although *tumor necrosis factor* alone might be adequate in a different context.

2) rabies immunoglobulin (RIG)
3) IGG receptor
4) Tumor necrosis factor 1

Application of the semantic constraint can result in apparent inconsistencies in the corpus (*immunoglobulin* is sufficient on its own in some sentences in the corpus, but is insufficient in (2)). However, we believe it is important that the tagged entity retain its true meaning in the *sentence context*.

Tokenization problems in GENETAG have been addressed in GENETAG-05 by switching to character-based indices. This helps make the corpus more robust, but it introduces embedded gene/protein entities that were previously untagged (*Ras*-mediated transformation, *PHA*-stimulated peripheral blood lymphocytes).

4 MedPost Review and Additions

The MedPost corpus (Smith et. al., 2004) originally contained 5 700 tokenized sentences. An additional 1 000 annotated sentences have been added for this release. Each sentence in the MedPost corpus is fully tokenized, that is, divided into non-overlapping annotated portions, and each token is annotated with one of 60 part of speech tags. Minor corrections to the annotations have been made since the original release.

Since most of the original corpus, and all of the sentences used for training the MedPost tagger, were in the area of molecular biology, we added an additional 1 000 sentences selected from random MEDLINE abstracts on the subject of clinical medicine. As a preliminary result, the trained MedPost tagger achieves approximately 96.9% accuracy, which is comparable to the 97.4% accuracy achieved on the subset of 1 000 sentences selected randomly from all of MEDLINE. An example of a sentence from the clinical medicine collection is

Evidence$_{NN}$ is$_{VBZ}$ now$_{RR}$ available$_{JJ}$ to$_{TO}$ show$_{VVI}$ a$_{DD}$ beneficial$_{JJ}$ effect$_{NN}$ of$_{II}$ bezafibrate$_{NN}$ on$_{II}$ retarding$_{VVGN}$ therosclerotic$_{JJ}$ processes$_{NNS}$ and$_{CC}$ in$_{II}$ reducing$_{VVGN}$ risk$_{NN}$ of$_{II}$ coronary$_{JJ}$ heart$_{NN}$ disease$_{NN}$.

In addition to the token-level annotations, all of the gerunds in the MedPost corpus (these are tagged *VVGN*) were also examined and it was noted whether the gerund had an explicit subject, direct object, or adjective complement. This annotation is stored with an annotation of type *gerund.* To illustrate, the two gerunds in the previous example, *retarding* and *reducing* both have direct objects (*retarding processes* and *reducing risk*), and the gerund tag is entered as "o". The gerund annotations have been used to improve a noun phrase bracketer able to recognize gerundive phrases.

References

Tanabe, L and Wilbur, WJ. 2002. Bioinformatics, 18, 1124-1132.

Tanabe L, Xie N, Thom, LH, Matten W, Wilbur, WJ: GENETAG: a tagged gene corpus for gene/protein named entity recognition. BMC Bioinformatics 2005.

Smith, L, Rindflesch, T, and Wilbur, WJ. 2004. MedPost: a Part of Speech Tagger for Biomedical Text. Bioinformatics, 20(13) 2320-2321.

Yeh A, Hirschman L, Morgan A, Colosimo M: BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics 2005.

Entity Type	Problem	GENETAG-05 Convention	Positive Examples	Negative Examples
Protein Families	Some are named after structural motifs.	Do not tag structures alone, but tag structurally related gene and protein families.	*Zinc finger protein, bZIP transcription factor, homeobox gene, TATA binding protein*	*Zinc finger, helix-turn-helix motif, leucine zipper, homeobox, TATA box*
Domains	Name can refer to 1) the amino acid content of a sequence (*PEST*), 2) the protein that binds the sequence (*TFIIIA DNA binding domain*), 3) a homologous gene (*SH2 - Src homology domain 2*), 4) the first proteins in which the domain was discovered (*LIM, PDZ*), or 5) structural entities (*POZ, zinc finger domain*).	Tag only if the domain refers to a gene or protein. Immuno-globulin regions are tagged. (*VH* refers to the *Immuno-globulin heavy chain V region*).	*BTB domain, LIM domain, HECT domain, VH domain, SH2 domain, TFIIIA DNA binding domain, Krüppel-associated box (KRAB) domains, NF-IL6 beta leucine zipper domain*	*PEST domain, SR domain, zinc finger domain, b-Zip domain, POZ domain, GATA domain, RS domain, GAR domain*
Boxes, Response Elements and Sites	Name can refer to 1) the sequence or site itself (*TAAG*), 2) a non-protein that binds to it (*Glucocorticoid Response Element*), 3) a protein that binds to it (*Sp1*), or 4) to homologous genes (*VL30*).	Tag only if the sequence or site refers to a gene or protein.	*VL30 element, Zta response elements, activating protein 1 (AP-1) site, Ets binding site, SP1 site, AP-2 box*	*GRE, TRE, cyclic AMP response element (CRE), TAAG sites, TGn motif, TAR element, UP element*
Hormones	Some are peptide hormones.	Tag only peptide hormones.	*Insulin, Glucagon, growth hormone*	*Estrogen, Progesterone, thyroid hormone*
"and" constructs	Some conjuncts require the entire construct.	Unless both conjuncts can stand alone, tag them together.	*TCR alpha and beta, D-lactate and D-glycerate dehydrogenase*	*TCR alpha, beta, D-lactate, D-glycerate dehydrogenase*
Viral Sequences	Promoters, enhancers, repeats are distinguished by organism.	Tag only if the organism is present.	*Viral LTR, HIV long terminal repeat, SV40 promoter*	*LTR, long terminal repeat*
Sequences	Some sequences lack gene or protein names.	Tag only if a gene name is included.	*NF kappa B enhancer (TGGAAATTCC)*	*TCTTAT, TTGGGG repeats*
Embedded Names	Some names are embedded in non-gene text.	Tag only the gene part.	*P-47*-deficient, *ras*-transformed	*P-47-deficient, ras-transformed*
Transposons, Satellites	Often repetitive sequences.	Tag if specific.	*L1 element, TN44, copia retrotransposon*	*non-LTR retrotransposon*
Antibodies	Often use organism or disease name.	Tag if specific.	*anti-SF group rickettsiae (SFGR)*	*antinuclear antibody*
Alternative Transcripts	Names differ from primary transcript.	Tag if primary transcript named.	*I kappa B gamma,VEGF20*	*Exon 2, IIA*

Table 1: Some problematic gene/protein annotations and conventions followed in GENETAG-05.