

Improving Biomedical Corpus Annotation Guidelines

Zhiyong Lu^{*1}, Michael Bada¹, Philip V. Ogren¹, K. Bretonnel Cohen¹ and Lawrence Hunter¹

¹Center for Computational Pharmacology, School of Medicine, University of Colorado, Aurora, CO, 80045, USA

Email: Zhiyong Lu* - Zhiyong.Lu@gmail.com; Michael Bada - Mike.Bada@uchsc.edu; Philip Ogren - Philip.Ogren@uchsc.edu; K. Bretonnel Cohen - Kevin.Cohen@gmail.com; Lawrence Hunter - Larry.Hunter@uchsc.edu;

*Corresponding author

Abstract

Background: We annotated over a thousand GeneRIFs with respect to different biological entities (e.g., protein) and processes (e.g., protein transport) in order to develop and evaluate text mining methods in molecular biology.

Results: We monitored inter-annotator agreements (IAAs) between two human annotators on a weekly basis and found that IAAs had a significant improvement over the first 10-week annotation period. In this paper, we present in detail our span selection guidelines along with other useful annotation experiences.

Conclusions: Clear annotation guidelines play a critical role in high quality corpus annotation. The guidelines presented in this paper are designed to be general purpose and easy to follow. Complete guidelines are publicly available at <http://compbio.uchsc.edu/grifs/transport/guidelines>

Background

Clear annotation guidelines are important for achieving consistent high quality annotations of textual corpora [1–3]. However, for the biomedical domain, there are few published annotation guidelines [1, 4–6], and these mostly concern only gene/protein names [1, 4, 5]. Few published corpora provide annotation beyond this level, most notably GENIA [7] and BioIE [8].

In support of a project that has annotated over a thousand GeneRIFs (Gene Reference Into Function) with 31 semantic classes in the protein transport domain, we have produced a novel set of guidelines for span selection with four advantages over other published guidelines: (1) They are generally applicable throughout molecular biology. Originally developed for annotating biological entities pertinent to protein transport (e.g., transport mechanisms, locations), their generality is validated by the application of the same set of rules to annotate many other concepts in

gene expression; (2) They can be learned and applied well by annotators; (3) They generally require only simple, straightforward judgements; and (4) Some of these rules can be applied to produce data for new tasks. For example, the rules for the embedded entities and appositives are useful for NLP tool development (see Discussion for details).

Methods

Annotation process

We hired two annotators with advanced degrees in molecular biology. They each had 20 hours of training that included learning Knowtator [9], a general purpose text annotation tool that was developed as a Protégé plug-in [10] and thus takes advantage of its knowledge representation capabilities. After annotation commenced, biweekly meetings were conducted with the annotators to raise concerns and discuss problems, which often led to refinements in the annotation guidelines.

Annotation guidelines for span selection

Every new annotation starts with a word in a GeneRIF that best corresponds to a class in an ontology. We call this special word the *anchor word*. Typically, the anchor word is the base noun in a noun phrase (e.g., the word “receptor” in the phrase “nuclear estrogen receptor” [PMID: 9522357] refers to **protein**). (Words and phrases in *Courier* denote classes of the ontology). And the noun phrases here do not include any prepositional phrases. Less frequently, another part of speech will fill the role of anchor word, for instance: (i) adjectives (e.g., “nuclear” in “nuclear estrogen receptor” refers to **nucleus**); (ii) modifying nouns (e.g., “estrogen” in “nuclear estrogen receptor” refers to **small molecule**); and (iii) verbs (e.g., “translocates” in “IMP1 translocates to the nucleus” [PMID: 12921532] refers to **transport**). After identifying an anchor word, annotators follow the rules below:

1. Rule for modifiers: Include all preceding nouns or adjectives that modify the anchor word, as well as *trailing variant specifiers* (letters or numbers used to distinguish a specific entity from a more general one). For example, in “estrogen receptor alpha” [PMID: 16271083], “receptor” is the anchor word, and it has a preceding modifier “estrogen” and a trailing variant specifier “alpha”. According to this rule, all three words “estrogen receptor alpha” should be selected as one annotation. It is to be noted that we do not consider preceding articles (e.g., a, the), demonstratives (e.g., this, that), pronouns (e.g., its, they), and quantifiers (e.g., one, all) as modifiers; hence, they should not be included.

2. Rule for embedded entities: Annotate embedded entities separately. Because of the previous rule, an annotation can sometimes include one or more other entities. We call those entities *embedded entities*. For example, an annotation of a protein “estrogen receptor alpha” includes another entity “estrogen,” the ligand to which the receptor binds. In this case, we ask annotators to make a separate annotation for the embedded entity “estrogen.” Together with the previous rule, for “nuclear estrogen receptor,” three annotations should be made (underlined below), one for each anchor word. Note that when the anchor word is “estrogen,” the preceding adjective “nuclear” is not included in the text span because it modifies “receptor,” not “estrogen.”

nuclear estrogen receptor (anchor word “receptor”)
nuclear estrogen receptor (anchor word “estrogen”)
nuclear estrogen receptor (anchor word “nuclear”)

3. Rule for appositives: Make separate annotations for appositives. An appositive is a noun phrase that renames or describes another noun phrase. An appositive often appears immediately after its expansion, as in “estrogen receptor-alpha (ER-alpha)” [PMID: 14691461]. However, “Nur77” in the phrase “TR3/Nur77” [PMID: 14500374] is also appositive because “Nur77” and “TR3” are synonyms for the same entity (Entrez GeneID: 3164). Thus, according to our rule, both “ER-alpha” and “Nur77” should be annotated separately.

4. Rule for punctuation: Normally, punctuation will not be included in span selection. For example, the parentheses in “estrogen receptor-alpha (ER-alpha)” should never be part of an annotation span. However, punctuation that is a part of a word or name should be included in span. For instance, the connecting hyphen in “ER-alpha” should be included in an annotation.

5. Rule for conjunctions: Mark up each constituent of the conjunctions. The conjunction words “and”, “or”, “&” and sometimes “/” and “-” create interesting annotation challenges such as “estrogen receptor alpha and beta heterodimers” [PMID: 15803276]. In this case, a total of four separate annotations should be made (underlined below), which correspond to **molecular complex**, **protein**, **protein**, and **small molecule**, respectively.

estrogen receptor alpha and beta heterodimers
estrogen receptor alpha and beta heterodimers
estrogen receptor alpha and beta heterodimers
estrogen receptor alpha and beta heterodimers

Results and Discussion

IAAs over time

We compared annotations between two human annotators by using inter-annotator agreement (IAA):

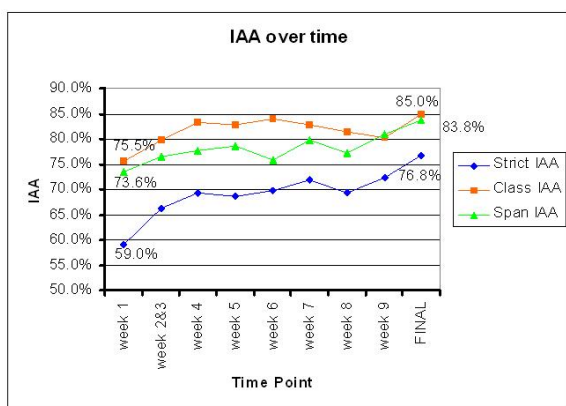
$$IAA = \frac{M}{M + NM}$$

where M is the number of annotations considered a match and NM is the number of non-matches. The sum of M and NM equals the total number of annotations. Three different IAAs were measured by changing the match criteria: *Strict IAA* requires that the annotations have the same class and span. *Type IAA* requires that two annotations have the same class but relaxes the span matching criteria such that the annotations’ spans need only overlap (rather than match exactly). *Span IAA* requires that two anno-

tations have only exact matching spans but does not compare their classes.

We computed the three IAAs on a weekly basis (in Figure 1) for the first ten weeks and found that the final IAAs are approximately 10% to 18% higher than the initial IAAs. There are multiple reasons for such increases, one of which we believe is due to the refinements in annotation guidelines. Other possibilities include that annotators became more familiar with the task, and that annotators followed the guidelines more strictly.

Figure 1 - IAA improvement over time



Rules can be well learned and applied

To investigate how well annotators use the rules after a period with relatively small or no rule changes (vs. constant changes during the first ten weeks), we examined the spans of 569 annotations that were recently marked up for 188 GeneRIFs. Table 1 shows the number of usages of each individual rule described in the previous section. As can be seen, approximately 40% (249/569) of the annotations required the annotator to apply one or more span selection rules (e.g., “fibroblast growth factor receptor 3 (FGFR3)” [PMID: 11731410]). We only found 10 cases where the annotator failed to follow the rules.

Table 1: Usage of span selection rules. Rules I – V corresponds to the rules for modifiers, embedded entities, appositive acronyms, punctuation, and conjunctions, respectively.

Span Rules	I	II	III	IV	V	Total
Correct Usages	92	43	14	96	4	249
Errors	2	6	2	0	0	10

Guideline design rationale

The design of span selection rules is based on two main criteria to guarantee the rules are: (1) general and applicable to different entities; and (2) capable of letting annotators produce consistent annotations.

The first criterion makes sure that when there is an ontological change (e.g., the addition of new concepts or even an entirely different ontology), the same rule can be applied. The second criterion makes sure the rules are practically useful to achieve high IAAs. An example is the rule for preceding modifiers, where annotators need only make simple, straightforward decisions regarding whether or not to include preceding modifiers, thereby facilitating consistent annotations.

Comparison to other guidelines

Our guidelines are comparable to those used in other projects but are different in that they: (i) are more explicit and consistent; (ii) require only simple, straightforward judgements; and (iii) can be applied to produce data for many new NLP tasks, such as recognizing nested named entities and identifying abbreviation definitions.

We will use the following examples for comparison of our guidelines to other published guidelines:

1. classII-positive B cell
2. IL-2 receptor
3. P-glyconprotein (P-gp)-related compounds

With regard to the first example, the entire phrase “classII-positive B cell” would be selected according to our rule for modifiers. The same annotation would also be made in the case of GENIA following a policy of “more specific concepts” [11], which is similar to our rule for modifiers because preceding modifiers typically lead to more specific concepts. However, our rule is more explicit. Furthermore, it rarely relies on annotators’ subjective judgement on what “more specific” means in various conditions, which according to [7] “may seem arbitrary”.

In the second example, a single annotation for “IL-2 receptor” but not “IL-2” would be made in GENIA and GENETAG [1] according to similar policies “mentioned substance only” [11]. Following our rule for embedded entities, we would make two separate annotations, “IL-2 receptor” and “IL-2”, because they refer to two distinct entities. This is desirable for two reasons. First, separate annotations are useful for developing and evaluating au-

tomatic methods of recognizing nested named entities [12]. Second, separate annotations are useful for more complex knowledge representations. For example, if we decide to add a slot (attribute) into the class `protein` to formally represent the ligand a protein binds to, then a separate annotation “IL-2” will be required to fill the slot. Finally, if some tasks (e.g., BioCreAtIve 1A [13]) require only the longer expression “IL-2 receptor”, we could easily remove “IL-2” during post-processing because we could determine “IL-2” is embedded in the longer expression “IL-2 receptor” by examining their spans. However, it is not a straightforward procedure to programmatically extract a nested entity from a single, undivided text span. For example, it is difficult to retrieve embedded entities such as “human lung” (which refers to `organ`) from a longer expression “human lung epithelial cell” (which refers to `cell`).

In the third example, a separate annotation for “P-gp” would not be made in BioIE [6] because it is nested in a longer term (“P-glycoprotein (P-gp)-related compounds”). However, the BioIE guidelines, in general, do require annotators to mark up abbreviations separately (e.g., annotate “GIST” separately in “gastrointestinal stromal tumor (GIST)”). In contrast, our rule for appositives requires annotators to always mark up abbreviations, which resulted in more consistent annotations that can be useful for developing and/or evaluating automatic tools like [14]. Furthermore, since our rule has no exceptions, it helps to reduce the total number of the rules, thus making it less difficult to be learned.

Other helpful experiences

In addition to detailed span selection guidelines, we found it is important to provide concrete examples, especially for classes that are sometimes hard to differentiate from one another (e.g., `small molecule` vs. its parent `molecule`).

Another useful experience is to decompose the complex annotation project into coherent subparts (Martha Palmer, personal communication), each of which can then be focused on individually. We switched to this approach by first asking the annotators to only mark up mentions of `cellular component` and its subclasses in all GeneRIFs, then mentions of `protein` in a second pass, and finally mentions of `protein transport` in a third pass.

Conclusions

We have described several notable features in our span selection guidelines by focusing on noun phrases because their annotation is usually more

difficult and thus requires more attention. Guidelines for others (e.g., verb phrases) can be found at the paper supplementary website. We conclude that our guidelines have advantages over other published guidelines and can be useful for many other similar annotation projects.

Acknowledgements

This work was supported by NIH grant R01-LM00811 (LH). We thank Sue Brozowski, Lynne Fox, and Manuel Miranda. We thank people from BioIE, GENETAG, and iProLink for making their guidelines publicly available.

References

1. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ: **GENETAG: a tagged corpus for gene/protein named entity recognition**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
2. Colosimo ME, Morgan AA, Yeh AS, Colombe JB, Hirschman L: **Data preparation and interannotator agreement: BioCreAtIve task 1B**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
3. Blaschke C, Leon EA, Krallinger M, Valencia A: **Evaluation of BioCreAtIve assessment of task 2**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
4. Inderjeet M, Zhangzhi H, Bae JS, Ken S, Matthew K, Jon Pa: **Protein name tagging guidelines: lessons learned**. *Comparative and Functional Genomics* 2005, **6(1-2):72-76**.
5. Vlachos A, Gasperin C, Lewin I, Briscoe T: **Bootstrapping the recognition and anaphoric linking of named entities in drosophila articles**. In *Proceedings of Pacific Symposium on Biocomputing* 2006:100-111.
6. **BioIE online annotation guidelines** [http://bioie.ldc.upenn.edu/wiki/index.php/Main_Page].
7. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus—semantically annotated corpus for bio-textmining**. *Bioinformatics* 2003, **19 Suppl 1**.
8. Seth K, Bies A, Liberman M, Mandel M, Mcdonald R, Palmer M, Schein A: **Integrated annotation for biomedical information extraction** 2004.
9. Ogren P: **Knowtator: A Protege plug-in for annotated corpus construction**. In *Proceedings of HLT-NAACL 2006 Demonstrations, NY, NY* 2006.
10. Noy NF, Sintek M, Decker S, Crubezy M, Ferguson RW, Musen MA: **Creating Semantic Web Contents with Protege-2000**. *IEEE Intelligent Systems* 2001, **2(16):60-71**.
11. Ananiadou S, Mcnaught J: **Corpus Annotation in Biology**. In *Text Mining for Biology And Biomedicine*, Artech House Publishers 2005:188-211.
12. Gu B: **Recognizing nested named entities in GENIA corpus [abstract]**. In *Proceedings of BioNLP workshop of HLT-NAACL 2006, Brooklyn, NY* 2006.
13. Yeh A, Morgan A, Colosimo M, Hirschman L: **BioCreAtIve task 1A: gene mention finding evaluation**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
14. Schwartz A, Hearst M: **A Simple Algorithm For Identifying Abbreviation Definitions in BioMedical Text**. In *Proceedings of PSB 2003, Lihue, Hawaii* 2003.