Using the Gene Ontology to Annotate Biomedical Journal Articles

Michael Bada and Lawrence Hunter
University of Colorado Denver, Aurora, CO, USA

Abstract

We are creating a gold-standard corpus of manually annotated full-text biomedical journal articles toward natural-language-processing applications. Central to this is our use of entire ontologies of the Open Biomedical Ontologies initiative as well as other terminologies as term sources, in contrast to most other such annotation projects, which have used small, ad hoc schemas. In addition to the standard difficulties in such annotation projects, each of the terminologies we have used has idiosyncrasies and ambiguities that present further challenges to consistent, high-quality annotation of these articles. In this paper we present and discuss the most salient of these with regard to the Gene Ontology that we have encountered and addressed in our annotation guidelines and training. The utility of these guidelines can be seen in the high and still-increasing interannotator-agreement statistics that we continually monitor.

Introduction

Gold-standard annotated biomedical corpora are essential for the training and evaluation of advanced biomedical natural-language-processing (NLP) systems, as evidenced by the significantly improved performance statistics of such systems trained on relevant corpora[1,2]. We have therefore embarked on an ambitious project to create a manually annotated corpus of full-text biomedical journal articles as a gold-standard community resource, which we are calling the CRAFT (Colorado Richly Annotated Full-Text) Corpus. In addition to manually annotating full-text papers, we are using all terms of select ontologies of the Open Biomedical Ontologies (OBO) initiative[3] (as well as other terminologies) as the annotating term sources, the first such effort of which we are aware. One part of this effort is annotation of these articles using the entire Gene Ontology (GO), which is comprised of terms representing biological processes (BP), molecular functions (MF), and cellular components (CC)[4].

In addition to the standard difficulties in this type of annotation project, our effort is particularly challenging due to our annotating full-text articles and using the full term sets of OBOs. Each of the terminologies we have used has its own idiosyncrasies and ambiguities that present further challenges to consistent, high-quality annotation and thus to text mining of these articles. In this paper we present and discuss the most salient of these issues that we have encountered in using the GO and addressed in our annotation guidelines and training. The utility of these guidelines can be seen in the high interannotator-agreement (IAA) statistics that we continually monitor.

Methods

Manual annotation is performed in an annotation tool developed within our lab called Knowtator[5], which is implemented as a plugin to Protege-Frames[6]. In an effort to reduce the annotators' workloads, the articles were preprocessed by automatically tagging them with terms from the relevant ontology; the annotators then check these annotations, making any needed deletions or corrections, and mark up anything else that was missed by the preprocessing. All of the annotators' work is further reviewed by the project lead (MB), with subsequent corrections being made. All tagging is saved as standoff annotation. IAA was calculated by comparing one annotator's set of annotations with the set of annotations created as a result of the project lead's review of the set, and Knowtator was used to calculate IAA statistics. Though the GO is continually evolving and growing, we are using a single static version (dated November 20, 2007, when this project was initiated) that contains 14,306 biological processes, 7,984 molecular functions, and 2,047 cellular components.

Guidelines for Applying the GO toward NLP Annotation

Given that we are using an ontology of more than 24,000 terms to annotate biomedical journal articles and that we strive for high IAA statistics, clear, well-conceived annotation guidelines are critical. Here we present the higher-level issues we have encountered and how we have addressed them in our guidelines. (We also have lower-level, linguistics-based guidelines, but this is outside the scope of this paper.)

1. Rules for General Biological Processes

We have found that general words indicating biological processes are particularly difficult to consistently annotate. For example:

(1) Bicarbonate formation is important for aqueous humor secretion from the ciliary processes and carbonic anhydrase (CA) facilitates this secretion. [PMID:11532192]

(2) Cells lacking BRCA1/2 fail to form damage-induced subnuclear RAD51 foci with normal efficiency, suggesting that these proteins are required for the formation of recombinase complexes at the sites of DNA damage. [PMID:11597317]

In (1), "formation" is closest to the biological-process term `biosynthetic process` (which encompasses the building up of more complex molecules from simpler ones), while in (2), "formation" is closest to the term `cellular component assembly`, which includes assembly of protein complexes, as is the case here. Rules have helped greatly in the annotation of such general process words. For example, pertaining to the above examples, if lexical variants of "form", "create", "assemble", etc. are applied to molecules, it is likely a biosynthetic process; if it is applied to cellular components, it is likely a cellular-component assembly. Annotating words denoting formation of cells and higher-level anatomical structures consistently is still difficult, as the GO cell and anatomical-structure development terms are arranged in a complicated (and some would say unintuitive) way. We plan on submitting our suggestions to the GO curation team toward making this part of the ontology clearer.

2. Molecular Functions vs. Biological Processes

The GO MF subontology is in a somewhat ambiguous state. The overwhelming majority of the MF terms are defined as molecular-level processes. (For example, the definition of `binding` is the "selective, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule".) Using the terminology of the Basic Formal Ontology (BFO), an upper-level ontology that is being used by members of the OBO Consortium toward evolution of the OBOs[7], a molecular function in this view is an occurrent, *i.e.*, a process, at the molecular level. However, there are those within the OBO Consortium who assert that a GO molecular-function term denotes not a process but a proper function, which is a dependent continuant, *i.e.*, an abstract entity that depends on the existence of another entity—essentially, the potential or functionality inherent in an entity to have a process realized[3]. It appears that the MF subontology will eventually evolve to this latter conceptualization.

This is relevant to our annotation work because this dichotomy between a process and the functionality to effect a process will often be expressed differently in natural-language text and thus should be annotated differently. (However, passages are often ambiguous with regard to process or function.) We have taken the view that molecular functions are processes, partly because this is how they are mostly currently defined and partly because mentions of processes are more common and more straightforwardly indicated in natural language than are functions to effect processes and thus more easily annotated.

A conflated issue is the fact that there are corresponding terms in the BP and MF subontologies that are extremely difficult to differentiate given a textual mention, even using their defintiions, *e.g.*, BP signal transduction and MF signal transducer activity, BP regulation of transcription and MF transcription regulator activity, BP caspase activation and MF caspase activator activity, as well as many corresponding BP transport and MF transporter activity terms. We have mostly dealt with this thus far by using most of the MF terms only to annotate text matching the term itself or an exact synonym (*e.g.*, kinase activity), an acceptably close synonym (*e.g.*, kinase functionality), or the corresponding continuant (*e.g.*, kinase, as discussed later in the paper). This is a suboptimal solution, but using the MF terms in this restricted way has allowed us to maintain our high levels of interannotator agreement.

A merging of the current MF subontology into the BP subontology to create one ontology of occurrents, from molecular-level to organism-level, would go a long way in ameliorating these two interrelated issues. We realize that this may seem radical, but it has been considered before, *e.g.*, at the 2008 NCBO Relation Ontology Expert Meeting[8]. We assert that this should involve merging the corresponding BP & MF term pairs such as those aforementioned. While the BP ontology would be an ontology of biological occurrents, the MF ontology could be redefined as an ontology of biological functions. An ontology of functions could likely be managed semiautomatically, as it would mostly mirror the relevant portions of the process ontology.

3. Noncanonical, Pathological, and *Ex Vivo* Entities and Processes

The GO is charged with representing canonical biological processes, molecular functions, and cellular components, and so we attempted at first to limit annotation to such canonical entities and processes. However, this turns out to be a deceptively difficult task. Sometimes the noncanonicality or pathology is explicit, as in:

(3) There were enlarged extracellular spaces between cells in the equatorial/bow region in 5 wk old alphaA/BKO lenses. [PMID:12546709]

In (3), it is obvious that the extracellular spaces are noncanonical in that they are larger than normal. But many times the noncanonicality or pathology can only be inferred from a very careful reading and comprehension of the article, and many other times it is not at all clear. This is especially due to the fact that most biological articles involve experiments with organisms or components of organisms in which they are subjected to all sorts of procedures, substances, and environments they would not normally encounter. Our solution is to annotate all mentions of GO entities and processes, even those that are explicitly noncanonical or pathological (so long as all of the other rules are followed, of course). Thus, in (3), "extracellular spaces" is annotated with the GO CC term `extracellular space` even though they are noncanonical in terms of their sizes.

A related issue is, given that the GO is in the domain of naturally occurring *in vivo* processes nd cellular entities, whether or not their *ex vivo* counterparts should be annotated. Analogously, to maximize our IAA, we decided to annotate all such *ex vivo* entities and processes; thus, a binding is a binding whether it takes place in an organism or in a beaker.

Smith *et al.* have written of noncanonical anatomical parts, which, in their representation with the Ontology of Biomedical Reality, are siblings of canonical anatomical parts; both of these are subsumed by a superclass of anatomical structures[9]. We analogously are viewing the entities and processes of the GO as these more general concepts that encompass both canonical and noncanonical instances.

4. Verb Nominalizations as Occurrents

Verb nominalizations can refer to either occurrents or continuants; Simon and Smith have written of such duality of certain biomedical terms (*e.g.*, dilation, dislocation) and how they address it in their LinKBase system[10]. As an example from our corpus:

(4) The vesicle formation goes along with several other changes in the red blood cell like cytoskeleton rearrangements and changes in the phospholipid orientation in the cellular membrane. [PMID:12925238]

In (4), "formation" clearly refers to a process, while "orientation" is a dependent continuant in that it is an attribute of the cellular membrane. We instruct our annotators to not annotate relevant mentions of such words if they clearly denote continuants since GO biological processes are occurrents. However, it is sometimes ambiguous whether such a relevant mention refers to either the occurrent or to the dependent continuant, *e.g.*, "distribution" in:

(5) The differentiation and distribution of specific mature neurons was examined in our previous study at adult stages with the expression of striatal markers such as preproenkephalin and Gad65/67. [PMID:15882092]

In such a case where one of the possible readings denotes an occurrent, we instruct the annotator to mark up the mention. If we had an ontology of the corresponding dependent continuants, we would also mark up such a mention with the dependent continuant term, as we encourage multiple annotation to capture the ambiguity of the expression.

5. Continuants with Molecular Functions

Mentions of continuants that have functions that can be realized in processes are often more frequent than the processes themselves; this is especially true for the molecular-function terms. For example, mentions of recombinases are more frequent than mentions of the MF term recombinase activity. We wished to capture these lexically analogous mentions, so each such mention is annotated with the corresponding term and also with the class `continuant`, the general term denoting an entity in the BFO. Thus, each such mention is doubly annotated as a process and as an entity. This is actually not semantically correct, as something cannot be both a continuant and an occurrent according to the BFO since these are disjoint classes in the BFO. It would be better to annotate each such mention once, as a continuant that has the corresponding function, *i.e.*, by annotating as a continuant and then adding a restriction to it, but Knowtator is not currently capable of this type of representation. Nevertheless, our

methodology enables us to capture all information that can then be easily transformed into a more semantically correct representation in our annotation repository.

Results

At this stage of our annotation of our 97-article corpus, we have created 8,279 annotations of cellular components (which is completed) and 18,996 annotations of molecular functions and biological processes in 44 articles (which is ongoing). (One annotator marked up articles with GO CC terms and another is annotating with GO BP and MF terms.)

To demonstrate the utility of our guidelines, we present the IAAs (calculated approximately weekly) for our two GO annotation passes in Figure 1. The annotation of the cellular components quickly rose to approximately 90% or higher, while the annotation of the biological processes and molecular functions started very low (9.7%) but has significantly risen since, with the last few data points at approximately 80%. There are large oscillations in the graph that are partly due to the fact that this annotator is typically able to annotate only one or two articles per period. A given article often has many mentions of a relatively small set of GO terms, and the IAA statistics are subject to such variation if the two annotators consistently annotate these numerous mentions of this relatively small set of GO terms differently.

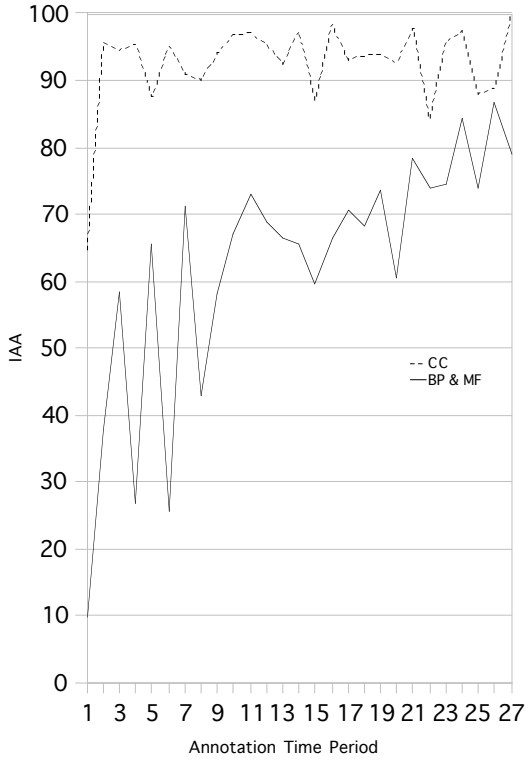

Figure 1. IAA statistics for the GO CC and BP/MF annotation projects over time. Each time period is approximately one week.

Conclusions

We have encountered issues in the course of our project to annotate biomedical journal articles with the whole of the GO, and we have attempted to address them with the guidelines presented in this article. These guidelines have in part allowed us to achieve high IAA statistics using the ontology as a very large annotation schema. We assert that these issues will also be troublesome for attempts at programmatic annotation and text mining of biomedical journal

articles, which many see as necessary in the future. We have presented several suggestions to the GO that we believe could ameliorate these issues.

Acknowledgements

This work is supported by NIH 5 T15 LM009451-02 and JDF 110200801921.

References

1. Tsuruoka Y, Tateishi JD, Ohta T, McNaught J, Ananiadou S, Tsujii J. Developing a robust part-of-speech tagger for biomedical text. Proc 10th Panhellenic Conf on Informatics. 2005; 382-392.
2. Lease M, Charniak E. Parsing Biomedical Literature. Natural Language Processing, Springer Berlin/Heidelberg. 2005; 58-69.
3. Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech. 2007;25:1251-1255.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25-29.
5. Ogren, PV. Knowtator: A plug-in for creating training and evaluation data sets for Biomedical Natural Language systems. Proc 9th Internat Protege Conf. 2006.
6. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubezy M, Eriksson H, Noy NF, Tu SW. The Evolution of Protege: An Environment for Knowledge-Based Systems Development. Internat J of Human-Comp Studies. 2003; 58(1):89-123.
7. Grenon P, Smith B, Goldberg L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. Ontologies in Medicine. IOS Press, Amsterdam. 2004;20-38.
8. http://bioontology.org/wiki/index.php/ OntologyRelations
9. Smith B, Kumar A, Ceusters W, Rosse C. On carcinomas and other pathological entities. Comp and Func Genom. 2005;6:379-387.
10. Simon J and Smith B. Using Philosophy to Improve the Coherence and Interoperability of Applications Ontologies: A Field Report on the Collaboration of IFOMIS and L&C. Proc 1st Workshop on Phil and Inform. 2004.