An Overview of the CRAFT Concept Annotation Guidelines

Michael Bada
Lawrence E. Hunter
University of Colorado Denver
Anschutz Medical Campus
Aurora, CO, USA
mike.bada@ucdenver.edu
larry.hunter@ucdenver.edu

Miriam Eckert
Martha Palmer
University of Colorado Boulder
Boulder, CO, USA
miriam_eckert@jdpa.com
martha.palmer@colorado.edu

Abstract

We present our concept-annotation guidelines for an large multi-institutional effort to create a gold-standard manually annotated corpus of full-text biomedical journal articles. We are semantically annotating these documents with the full term sets of eight large biomedical ontologies and controlled terminologies ranging from approximately 1,000 to millions of terms, and, using these guidelines, we have been able to perform this extremely challenging task with a high degree of interannotator agreement. The guidelines have been designed to be able to be used with any terminology employed to semantically annotate concept mentions in text and are available for external use.

1 Introduction

Manually annotated gold-standard corpora are becoming increasingly critical for the development of advanced NLP systems. At the same time, the use of ontologies as formal representations of domain-specific knowledge is being seen in a wide range of applications, particularly in the biomedical domain. We are synergistically creating a gold-standard corpus called the Colorado Richly Annotated Full-Text (CRAFT) Corpus that pushes the boundaries of both of these prominent types of resources. For this project, we are manually annotating a collection of 97 full-text biomedical journal articles comprising a total of more than 750,000 words, as opposed to the sentences or abstracts upon which other gold-standard corpora have focused. Additionally, while most other related corpora have used small annotation schemas consisting of a few to several dozen classes for their semantic annotation, we are employing the full sets of terms, ranging from approximately one thousand to several tens of thousands of terms, of select ontologies of the Open Biomedical Ontologies (OBO) Consortium, the most prominent set of biomedical ontologies (Smith *et al.*, 2007), as well as several other significant large biomedical controlled terminologies. The terms of these ontologies and terminologies, which serve as the classes of the semantic annotation schema for the this corpus, are continually under development by biomedical researchers and knowledge engineers and are widely used throughout the biomedical field, as opposed to other annotation schemas that are often idiosyncratic and not likely reusable for other tasks. Furthermore, though these ontologies have been used for a variety of NLP tasks, they have not been used in their entirety toward gold-standard markup of text.

With regard to the CRAFT Corpus project, we have previously written of desiderata in using large ontologies and terminologies for semantic annotation of natural-language documents (Bada and Hunter, 2009a) and of semantic issues in the use of one of the ontologies we are using, the Gene Ontology (Bada and Hunter, 2009b). In this paper, we present a brief overview of the concept[1] annotation guidelines we are using for this corpus and the motivations behind our choices. With these guidelines, our annotators have routinely achieved 90+% agreement with the project lead on all but the one most challenging terminological annotation passes, which currently is more than 80%. The guidelines were designed to be reusable regardless of the

[1] Throughout this document, "concept", "class", and "term" are used interchangeably.

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 207–211,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

ontology/terminology being used for semantic annotation, and we have indeed used them with minimal exceptions for concept annotation of our corpus using eight orthogonal large ontologies and terminologies.

2 Overview of the CRAFT Corpus

The CRAFT Corpus is a collection of 97 full-text biomedical journal articles that is being richly annotated both syntactically and semantically and is designed to be an open community resource for the development of advanced bioNLP systems. The 97 articles of the corpus comprise the intersection of articles that are open-access and that have been used as evidential sources for Gene Ontology (GO) annotations of genes and gene products of the laboratory mouse by our collaborators who serve as the official GO curators of the preeminent mouse database. (The GO, the flagship OBO, is an ontology composed of three subontologies representing the specific molecular functions (MF) of genes and gene products, the higher-level biological processes (BP) in which they participate, and the cellular components (CC) in which they localize (Ashburner *et al*., 2000). GO annotations, which are entirely different from the annotations we discuss in the work presented here, are created by labeling genes and gene products of organisms with GO terms.)

These articles in their entirety are being syntactically annotated by sentence segmentation, tokenization, part-of-speech tagging, and treebanking. The articles' nouns and noun phrases are also being coreferentially annotated (Cohen *et al*., 2010). Though these branches constitute a significant amount of the annotations of the project, they are outside the scope of this paper. Furthermore, we are working on creating assertional annotations between the concept annotations via relations.

Six ontologies of the OBO library and two additional controlled terminologies have thus far been selected for concept annotation of these articles on the bases that these are relatively well-constructed knowledge representations, are widely used by bioinformaticians and/or biomedical researchers, and represent concepts needed to extract significant biomedical assertions from the literature. In addition to the three aforementioned GO ontologies, the OBOs that were selected for concept annotation are the Cell Type Ontology (CL), which represents types of cells (Bard *et al*., 2005); the Chemical Entities of Biological Interest (ChEBI) ontology, which represents types of small molecules, parts of molecules, atoms, and subatomic particles (Degtyarenko *et al.*, 2008); and the Sequence Ontology (SO), which represents types of biological macromolecules and their components (Eilbeck *et al*., 2005). In addition to these ontologies, we are also annotating the articles with the terms of the NCBI Taxonomy, the most widely used Linnaean hierarchy of biological organisms, and the unique identifiers of the Entrez Gene database, the preeminent resource for species-specific genes (Sayers *et al*., 2009).

The annotation methodology, not presented here due to lack of space, has been presented in a previous publication (Bada and Hunter, 2009a).

3 Overview of the CRAFT Concept Annotation Guidelines

Concept annotation entails annotating text with concepts, *i.e.*, classes or terms from ontologies or terminologies. (We use this more expansive term as opposed to named-entity annotation since several of the terminologies we are using contain terms representing processes and functions, which are annotated just as terms representing entities are.) Every mention (including abbreviations and misspellings) of every *explicitly represented* concept of the ontology or terminology is annotated, and the text selected must be as semantically close as possible—essentially semantically equivalent—to the term with with which it is annotated. Thus (as shown later), a mention of platelets is semantically annotated with a term representing platelets as opposed to the more common case of annotating with a more general term (*e.g.*, representing cells) selected from a much smaller annotation schema.

For each concept annotation, any selected text span must be adjacent on each of its boundaries to an appropriate delimiter. A whitespace character most often serves as a delimiter:

> **Ex. 1.** localization: :of: :annexin: :A7: :in: :platelets: :and: :red: :blood: :cells [PMID:12925238[2]]

(Colons indicate possible boundaries of annotations.) Any punctuation mark can also serve as a delimiter indicating a boundary of an annotation:

[2] For each example, the PubMed ID of the biomedical article from which it is extracted is shown.

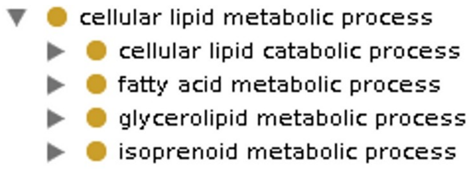
Ex. 2. To examine this:,: we analyzed the ability of red blood cells derived from the annexin A7 mice :(:*anxA7*:-:/:-:): to form exovesicles:.: [PMID:12925238]

Finally, beginnings and ends of documents can serve as boundaries of annotations.

It is important to note that letters (including non-Latin letters) and numbers can never serve as delimiters. Practically, this means that an annotation text span can never begin or end in between two letters, between two numbers, or between a number and a letter. These delimiters were chosen so that the annotator would not be burdened with the very difficult and time-consuming task of having to figure out what every letter of every abbreviation represented and whether they should be annotated; similarly, this avoids evaluation of any arbitrary part of any word (*e.g.*, whether the "cyto" of "cytological" should be annotated with the term `cell`[3]). This choice of delimiters sometimes prevents the annotator from creating an annotation that he or she may wish to create, but in our experience, this is a relatively rare occurrence, and it is a small price to pay for greatly simplifying an already extremely large and difficult task. Furthermore, it is a straighforward rule for both human and computational annotators to follow.

One primary motivation behind our strategy of annotating only explicitly represented concepts is the capture of the exact semantics of textual mentions; conversely, annotating a textual mention with a more general term (*e.g.*, annotating "platelet" with `cell`) entails loss of knowledge. A second motivation is that of making this task of semantic annotation doable: The alternative of annotating every mention of the concepts within the domain of a given terminology including all concepts within the domain that are not explicitly represented in the terminology rapidly becomes an overwhelming task with even a moderately sized terminology. For example, using this alternative strategy to annotate all mentions of ChEBI chemical concepts explicitly represented or not, if an annotator came across a mention of a chemical not represented in the ontology, *e.g.*, iodixanol, assuming he were not intimately familiar with the structure and function of iodixanol, he would have to first research this. From among the thousands of structural terms, he would have to annotate this mention with all relevant terms pertaining to its structure such as `amides`, `polyols`, `aromatic compounds`, and `organoiodine compounds` since this compound contains the corresponding chemical groups that define these types of molecules (and none of these terms subsumes another). Furthermore, he would have to evaluate annotating with all relevant terms from among the hundreds of ChEBI functional terms (*e.g.*, `xenobiotic`, `base`, `chromophore`, `cofactor`). This enormous amount of work becomes even more difficult when working with concepts that are not as precisely defined as, for example, the chemical structure terms.

Text spans that can be considered for annotation are dictated by syntax, and the text that is selected must be semantically equivalent to a term in the ontology/terminology. For example, for a noun, any modifying adjective or prepositional phrase can be considered for inclusion in the annotation if its inclusion results in a semantic match to a concept in the ontology/terminology.



▼ ● cellular lipid metabolic process
- ▶ ● cellular lipid catabolic process
- ▶ ● fatty acid metabolic process
- ▶ ● glycerolipid metabolic process
- ▶ ● isoprenoid metabolic process

Fig. 1. Part of the GO BP `cellular lipid metabolic process` hierarchy.

Ex. 3: Skeletal muscle is a major site to regulate whole-body fatty-acid and glucose metabolism. [PMID:15328533]

In Ex. 3, "metabolism" along with its premodifying "fatty-acid" (but not with its premodifying "whole-body") are selected for one annotation, as this is a semantic match to the GO term `fatty acid metabolic process`. Determiners and quantifiers are never included in concept annotation. Note that this is an example of a *discontinuous annotation*—an annotation consisting of two or more discontinuous spans of text, which is unambiguously represented as standoff.

The use of one or more terminologies in the semantic markup of text may result in overlapping and nesting annotations. *Overlapping* refers to the overlapping of the selected text of an annotation, in part or in whole, with the selected text of another annotation. *Nesting* is a type of overlapping in which the selected text of an annotation is a proper subset of the selected text of

[3] Names of ontological concepts are rendered in `fixed-width type` throughout this document.

another another. A nested annotation is created only if it is to be annotated with a term that is *not* a superclass of the term used in the nesting annotation. This is a trivial evaluation if the terms for the nesting and nested annotations are from different terminologies, as one cannot be a superclass of the other; if the terms are from the same terminology, one may or may not be a superclass of the other. There are no corresponding restrictions for overlapping annotations that are not nesting/nested annotations.

The full CRAFT Corpus annotation guidelines can be viewed at http://bionlp-corpora.sourceforge.net/CRAFT/CRAFT_concept_annotation_guidelines.pdf and are available for use by others under a specified Creative Commons license.

4 Results

To date, we have created more than 107,000 concept annotations; these are broken down by terminology in Table 1.

Terminology	# Annotations	# Articles
ChEBI	15,313	97
CL	8,290	97
Entrez Gene*	5,618	29
GO BP*	22,101	91
GO CC	7,247	97
GO MF*	5,563	91
NCBI Taxonomy	11,202	97
SO	32,502	97
Total	107,836	-

Table 1. Current counts of annotations and articles; * indicates an ongoing pass.

To illustrate the utility of our guidelines, we present the IAAs for six terminological passes of the corpus. As seen in Figs. 2 and 3, the annotators quickly reach and with few exceptions remain at a 90+% IAA level for all of the terminological passes except for the extremely challenging (and ongoing) GO BP & MF pass, currently at a typical 80-85%. As presented previously, most of these data points are single-blind statistics; however, as a control, a small number were annotated double-blind, including three articles annotated with the SO, which resulted in an IAA of 89.9%, compared with a single-blind IAA of 90.4% for the previous week, suggesting that these single-blind IAAs are unlikely to be significantly biased.

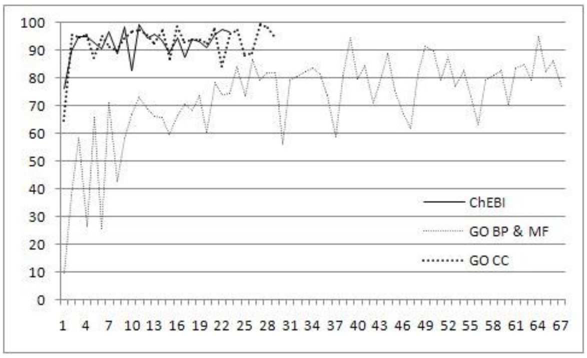


Fig. 2. IAA vs. number of training sessions for annotation of the corpus with ChEBI, GO BP & MF, and GO CC.

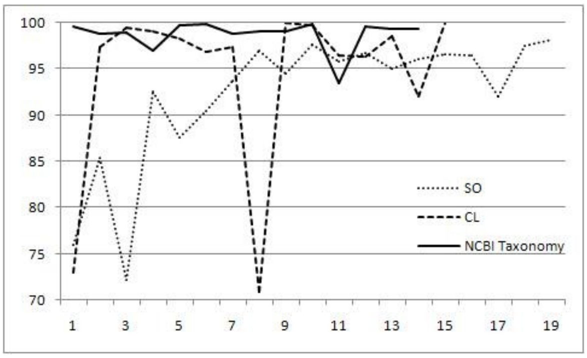


Fig. 3. IAA vs. number of training sessions for annotation of the corpus with SO, CL, and NCBI Taxonomy.

5 Conclusions

We have succinctly presented our concept-annotation guidelines, with which we routinely achieve high IAAs in the semantic annotation of full-text biomedical journal articles. The decisions behind these guidelines were made to maximally facilitate both manual and programmatic annotation of text with the full term sets of terminologies, particularly large ones. Foremost, the decision to annotate a part of the text with a term is based on whether this text is a direct semantic match to an explicitly represented term, and the specific selection of text is cleanly dictated by syntactic rules. Additionally, to greatly reduce the workload of our human annotators, a nested annotation is created only if the term to be used is not a superclass of the term used to annotate the nesting concept mention. These guidelines were designed to be used with any ontology or terminology and are available for others to use.

Acknowledgments

The authors gratefully acknowledge their support by NIH 5G08M009639 and 5T15 LM009451.

References

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. 2000. Gene Ontology: tool for the unification of biology. Nat Genetics, 25:25-29.

Bada, M. and Hunter, L. 2009a. Using Large Terminologies to Semantically Annotate Concept Mentions in Natural-Language Documents. Proceedings of the International Conference on Knowledge Capture (K-CAP) Semantic Authoring, Annotation and Knowledge Markup (SAAKM) Workshop 2009, Redondo Beach, CA, USA.

Bada, M. and Hunter, L. 2009b. Using the Gene Ontology to Annotate Biomedical Journal Articles. Proceedings of the International Conference on Biomedical Ontology (ICBO) 2009, Buffalo, NY, USA.

Bard, J., Rhee, S. Y., and Ashburner, M. 2005. An ontology for cell types. Genome Biology, 6(2), R21.

Cohen, K. B., Lanfranchi, A., Corvey, W., Baumgartner, Jr., W. A., Roeder, C., Ogren, P. V., Palmer, M., and Hunter, L. E. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. Proceedings of the 7th Language Resources and Evaluation Conference (LREC) Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), Valletta, Malta.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. 2008. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research, 36, Database Issue:D344-D350.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biology 6, R44.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvarov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 37, Database Issue:D5-15.