Gold-Standard Ontology-Based Annotation of Concepts in Biomedical Text in the CRAFT Corpus: Updates and Extensions

Michael Bada, Lawrence Hunter
University of Colorado
School of Medicine
Aurora, CO, USA
{mike.bada, larry.hunter}@ucdenver.edu

Nicole Vasilevsky, Melissa Haendel
Ontology Development Group, Library
Oregon Health & Science University
Portland, OR, USA
{vasilevs, haendel}@ohsu.edu

***Abstract*—Ontologies are increasingly used for semantic integration across disparate curated biomedical resources, while gold-standard annotated corpora are needed for accurate training and evaluation of text-mining tools. Bringing together the respective power of these, we created the Colorado Richly Annotated Full-Text (CRAFT) Corpus, a collection of full-length, open-access biomedical journal articles that have been manually annotated both syntactically and semantically with select Open Biomedical Ontologies (OBOs), the first release of which includes ~100,000 annotations of concepts mentioned in the text of 67 articles and mapped to the classes of eight prominent OBOs. Here we present our continuing work on the corpus, including updated versions of these annotations with newer versions of the ontologies, new annotations made with two additional OBOs, annotations made with newly created extension classes defined in terms of existing classes of the ontologies, and new annotations of roots of prefixed and suffixed words.**

Keywords—annotation, corpus, markup, ontology.

I. Introduction

With the ever-rising amount of biomedical literature, it is increasingly difficult for scientists to keep up with the published work in their fields of research, much less related ones. The use of natural language processing (NLP) tools can make the literature more accessible by aiding concept recognition and information extraction. As NLP-based approaches have been increasingly used for biocuration, so too have biomedical ontologies, whose use enables semantic integration across disparate curated resources, and millions of biomedical entities have been annotated with them. Particularly important are the Open Biomedical Ontologies (OBOs), a set of open, orthogonal, interoperable ontologies formally representing knowledge over a wide range of biology, medicine, and related disciplines [1].

Manually annotated document corpora have become critical gold-standard resources for the training and testing of biomedical NLP systems. This was the motivation for the creation of the Colorado Richly Annotated Full-Text (CRAFT) Corpus, a collection of 97 full-length, open-access journal articles from the biomedical literature [2, 3]. Within these articles, each mention of the concepts explicitly represented in eight prominent OBOs has been annotated, resulting in gold-standard ontology-based markup of genes and gene products, chemicals and molecular entities, biomacromolecular sequence features, cells and cellular and extracellular components and locations, organisms, biological processes and molecular functionalities. With these ~100,000 concept annotations among the ~800,000 words in the 67 articles of the 1.0 release, it is one of the largest gold-standard biomedical semantically annotated corpora. In addition to this substantial conceptual markup, the corpus is fully annotated along a number of syntactic and other axes, notably by sentence segmentation, tokenization, part-of-speech tagging, syntactic parsing, text formatting, and document sectioning.

In the several years since the initial release of the CRAFT Corpus, in addition to efforts within our group and in collaboration with others, including the first comprehensive gold-standard evaluation of prominent concept-recognition systems [4], it has already been used in multiple external projects to drive development of systems for biomedical curation, search, visualization, and semantic and syntactic NLP tasks (*e.g.*. [5, 6]). Here we present our continuing work on the corpus, including updated versions of these annotations with newer versions of the ontologies, new annotations made with two additional OBOs, annotations made with newly created extension classes defined in terms of existing classes of the ontologies, and new annotations of roots of prefixed and suffixed words.

II. Methods

All continuing work on the concept annotations of the CRAFT Corpus was performed in Knowtator, a plugin to Protégé-Frames [5]. (as was done for the v1.0 concept annotations). The lead annotator (MB) made updates to the v1.0 concept annotations using newer versions of the ontologies that had been used to mark up the articles by removing annotations of obsoleted classes, editing previously made annotations, and creating new annotations for new classes. A list of approximately 20 prefixes and suffixes was compiled, and roots of words with these affixes were annotated as their unaffixed analogs would be. As the

updating progressed with each ontology, corresponding extension classes were created to use for further annotation.

Annotation of the corpus with the Molecular Process Ontology (MOP) and Uberon was performed in one primary round (by NV) followed by a review (by MB) using the original concept annotation guidelines [6]. Roots of words with aforementioned affixes were also annotated, and extension classes were also created and used for additional annotation. The articles were annotated with a single ontology at a time and a batch at a time (8 articles per batch for the MOP and 4 articles per batch for Uberon), and interannotator agreement (IAA) was calculated for each batch using Knowtator's built-in IAA calculation functionality. The curators strove for IAA ≥ 90% for each annotation batch.

III. RESULTS AND DISCUSSION

So as to remain current and relevant, the v1.0 concept annotations of the corpus are being reviewed and updated by addition, editing, and deletion of annotations as appropriate, relying on newer versions of the eight OBOs previously used. Updating with four of these has been completed.

The extension of annotation of specific affixed root words is largely for consistency: In the v1.0 corpus, any whitespace or punctuation character could serve as an annotation delimiter; thus, "chromatin" of "anti-chromatin" would be annotated with the Gene Ontology class for chromatin (GO:0000785), but it could not be annotated within "antichromatin", as there is no delimiter. The rendering of such affixes is variable in that they can be nondelimited from their root words or delimited by whitespace or punctuation, so with this updating, the markup of such affixed words is now more consistent; furthermore, additional knowledge is captured. A specific list of such affixes to consider has been compiled and will be provided with the next release.

While creating the concept annotations for the v1.0 corpus, we encountered a variety of difficulties with annotating exclusively with explicitly represented OBO classes, including class ambiguity, lack of sufficiently generic classes, lack of classes for words consisting of combinations of multiple ontology classes, representation of the same concept in multiple ontologies and incompleteness of ontologies. To ameliorate these issues, we have been creating and using specific extension classes for concept annotations for the corpus update. All of these are formally defined in terms of explicitly represented OBO classes, and we intend to make these definitions available in OWL files in the next release. However, we also intend to release the annotations in sets both including and excluding these extension classes for users who respectively do and do not wish to make use of annotations with such classes in their work.

Finally, for the purpose of capturing additional types of biomedically relevant concepts, annotations have been created for the articles of the corpus using the classes of the MOP ontology of chemical processes [7] and the Uberon anatomical ontology [8]. Tables 1 and 2 display relevant statistics for the 67 articles of the public set, excluding and including use of extension classes, and IAA statistics are presented in Figure 1.

ontology	*total # annotations*	*average # annotations per article*	*median # annotations per article*	*max # annotations per article*
MOP	293 / 331	4 / 5	2 / 2	34 / 34
UBERON	12,238 / 15,051	183 / 225	130 / 169	578 / 709

ontology	*total # unique concepts*	*average # unique concepts per article*	*median # unique concepts per article*	*max # unique concepts per article*
MOP	19 / 20	2 / 2	1 / 1	6 / 6
UBERON	850 / 898	31 / 37	24 / 30	109 / 129

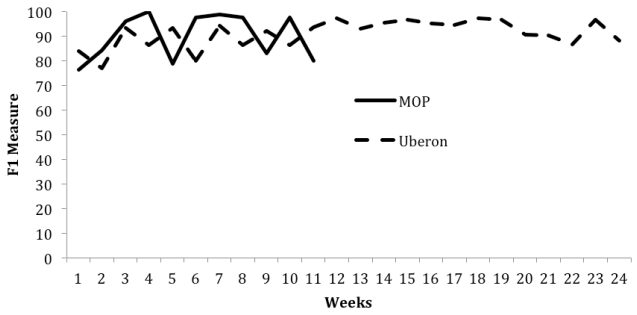

Figure 1: IAA (as F_1-measure) vs. annotation batch number.

IV. CONCLUSIONS

We have presented our continuing work on the gold-standard concept annotations of the CRAFT Corpus, including updated versions of the annotations with newer versions of ontologies, new annotations made with additional OBOs, annotations made with newly created ontology extension classes, and new annotations of roots of prefixed and suffixed words. We intend to soon release these updated annotations in future versions of the corpus, and we also have longer-term plans for further development of the corpus.

ACKNOWLEDGMENT

This work was supported by grant DARPA-BAA-14-14.

REFERENCES

[1] http://www.obofoundry.org

[2] Bada M, Eckert M, Evans D, Garcia K, Shipley K, et al. (2012) Concept Annotation in the CRAFT Corpus. BMC Bioinform 13:161.

[3] Verspoor K et al. (2012) A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinform 13:207.

[4] Funk C., Baumgartner W., Garcia B., Roeder C., Bada M. et al. (2014) Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bionform 15:59.

[5] Liu H et al. (2012) BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. J Biomed Semantics 3:3.

[6] Nunes T, Campos D et al. (2013) BeCAS: biomedical concept recognition services and visualization. Bioinform 29(15), 1915-1916.

[7] Ogren P.V. (2006) Knowtator: a Protégé plug-in for annotated corpus construction. Proc Hum Lang Tech Conf N Am Chap Assoc Comp Ling.

[8] Bada M et al. (2010) An overview of the CRAFT concept annotation guidelines. Proc 4th Ling Annot Wkshp, Assoc Comp Ling, 207-211.

[9] http://obofoundry.org/ontology/mop.html

[10] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE et al. (2011) Uberon, an integrative multi-species anatomy ontology. Genome Biol 13:R5.