

The Colorado Richly Annotated Full-Text (CRAFT) Corpus: A Resource for Biocurational Text-Mining Research

Michael Bada¹, Miriam Eckert², Kristin Garcia³, Donald Evans³, Dmitry Sitnikov⁴, William A. Baumgartner, Jr.¹, Philip V. Ogren⁵, Arrick Lanfranchi², Amanda Howard², William Corvey², Nianwen Xue², Kevin B. Cohen¹, Karin Verspoor¹, Judith A. Blake⁴, Martha Palmer², Lawrence Hunter¹

¹ University of Colorado Denver, Department of Pharmacology, Aurora, CO 80045, USA

² University of Colorado Boulder, Department of Linguistics, Boulder, CO 80309, USA

³ University of Colorado Boulder, Department of Molecular, Cellular, and Developmental Biology, Boulder, CO 80309, USA

⁴ The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

⁵ University of Colorado Boulder, Department of Computer Science, Boulder, CO 80309, USA

Research in various biomedical natural-language-processing tasks, notably concept identification and information extraction, more and more relies on well-annotated corpora as gold standards for training and evaluation. To this end, we have embarked upon the creation of the Colorado Richly Annotated Full-Text (CRAFT) Corpus as a resource for biocurational text-mining research. An initial corpus of 97 biomedical journal articles, primarily focused on the laboratory mouse, has been selected to be annotated in full (*i.e.*, all text but the bibliographic references contained within the articles) based on two criteria: (1) that they have been used as evidential sources in the annotation of genes or gene products with Gene Ontology (GO) terms, so that our annotations may be used to mine for this most prominent category of biocuration and (2) that they are freely available (*i.e.*, open-access), so as to permit unfettered distribution and research use. The annotation of this corpus is broadly divided into semantic and syntactic annotation, each of which includes a variety of annotation subtasks. As for the former, it is of significance that we are using community-accepted ontologies and terminologies, including those from the Open Biomedical Ontologies (OBO) Consortium, as the terms with which we are semantically annotating the texts. Furthermore, we are using these ontologies and terminologies in their entirety, as opposed to using subsets of terms; ours is the first such effort of which we know to undertake this. Thus far, we have finished annotating cells (using the OBO Cell Type Ontology) and cellular components (using the OBO GO cellular-component subontology), and we are currently annotating biological processes and molecular functions (using those subontologies of the OBO GO), chemicals, chemical groups, atoms, and subatomic particles (using the OBO ChEBI ontology), and organisms (using the NCBI organism taxonomy). We further intend to (at least) semantically annotate genes and gene products and biomedical sequences and also to relationally link these conceptual annotations. Syntactically, annotation tasks that we are performing (through a combination of automatic and manual techniques) include tokenization, part-of-speech tagging, and full parsing (*i.e.*, treebanking). Furthermore, we have recently begun coreferential annotation of these articles. All of our standoff annotations and the articles will be freely available, and we believe this will be an important resource providing ample avenues for research in (semi)automatic biocuration.