Positional Paper on a Semantic Web for Life Sciences

Alexander A. Morgan, Alexander S. Yeh, Marc Colosimo, Lynette Hirschman,
MITRE Corporation
Contact: amorgan@mitre.org

Our research primarily involves the application of natural language processing technology to biomedical literature in support of such applications as semi-automated functional annotation of proteins and genes, and gene name normalization for improved search and retrieval of text information. We have performed studies in the use of existing database resources in these efforts (Morgan, Hirschman et al. 2003) and together with CNB/CSIC-Madrid, we have organized and administered a challenge evaluation, BioCreAtIvE (Valencia, Blaschke et al. 2004), for text mining systems applied to biomedical literature. Our primary experience with ontologies is with GO (The Gene Ontology Consortium 2000), and with some of the specific hierarchical controlled vocabularies. These include the FlyBase Controlled Vocabulary (The FlyBase Consortium 1993) and the TVFac Hierarchy (http://www.tvfac.lanl.gov/right.html); our focus has been on automating the association of small excerpts of text and the underlying entities described (mentioned) in the text with concepts in the ontologies. We focus here on how existing ontologies and related resources can be augmented to aid text-mining and how text mining evaluation techniques can contribute to ontology evolution, by viewing ontologies as annotation guidelines when constructing/populating them.

Even as simple a task as determining which DNA sequence is being described when a gene is mentioned in a MEDLINE abstract can be challenging task. With an organism such as *Drosophila melanogaster* (with somewhat free-wheeling naming conventions), identifying the mentions of gene names can be non-trivial, given that *white, clock, dorsal,* and *period* are all gene names. Trying to associate a gene mention with a given functional code in GO is even more difficult, given the linguistic distance between a GO concept or description, and how this attribute is actually described in text (see the examples given below). We believe that these tasks can be facilitated by enriching the lexicon and using sets of synonyms from additional biological resources. If we can automate this process, this will make it far easier to link databases and to annotate records in the databases.

An ontology can be enriched by including synonymous descriptions of the concept that the node is intended to represent. A node in a generic biological ontology may include a unique identifier, a name, links out to parents and children, and sometimes a few sentences describing the concept. However, the name of the concept may be very different than anything that might appear in any text mentioning that concept or describing an object with its properties. A longer text description might or might not be helpful for these purposes. For example, an ontology of experimental techniques might include a concept such as *Immunoblot*, as does FlyBase. However, that term is unlikely to appear explicitly in the Material and Methods section; rather, we are more likely to see the descriptions of antibodies used and a description of the procedure. Another example is the GO code 0005388, *calcium-transporting ATPase activity*, which is unlikely to appear in a description of a protein associated with that code. However, GO includes

synonyms such as *calcium efflux ATPase, calcium pump,* and *sarcoplasmic reticulum ATPase* that might aid in recognition, particularly combined with a term dictionary that expands calcium to Ca2+, the way it most often appears in text.

The effort to develop a large number of LSID's (Life Science Identifier) should be a great help to text-mining efforts. Linking a GenBank accession number with another database with gene and protein annotations can help expand the synonymous variants and other key text fields that may be used. Also, mappings between ontologies can help deal with many of the previously mentioned issues, because short text descriptions in one ontology may be expanded in another.

Unfortunately, research is only just beginning on how to use the links between concepts in an ontology to improve text mining. Taking GO as an example, the correct annotation for a protein is the most specific (deep) functional annotation known. Although high up in the ontology, concept 0009987, *cellular process*, exists, and it would be accurate to annotate most proteins with this label, it is far too general to be relevant. Computer scientists have examined various distance measures, but an underlying problem is that graph distance may bear little relation to semantic distance in a human generated ontology. This sense of distance is important when trying to expand a match of terms or disambiguate the sense of text in a passage as it relates to an entry in the ontology.

The semantic distance is really encoded in how the ontology is used. Biological ontologies are used to label data, e.g. associate a GO code with a protein, label a patient record with an ICD-10 code, label a piece of data with an experimental method code, annotate the subject matter of a figure or graph, or link expression of a protein to a anatomical term in the FlyBase fly anatomy. When used in this way, ontologies tend to have a very skewed distribution of the labeling. (Lord et al 2003) proposes an information theoretic metric based on a posteriori distribution of genes annotated in the Gene Ontology. A set of labels with the highest information content would have a uniform distribution of the labels. Instead we see exponential decay curves, with a handful of concepts used repeatedly to label different instances, and many not used at all. Repeatedly used concepts may benefit from further refinement, e.g. added child concepts to provide more detailed information. The areas of the graph not visited at all show may show excess specificity. Of course, the skew also reflects an uneven advance of the state of knowledge, as well as trends in research, where certain areas receive more attention than others.

When developing an ontology, it is important to keep in mind that it is not only a representation of concepts and their relationships, but that it generally has specific uses. In the case of a biological ontology that is used to annotate biological 'entities', the concepts are directly associated with those entities, and it is important to make sure that the ontology can be used consistently by different annotators. If two individuals cannot consistently label the same entity with the same concepts from the ontology, then there is a problem with how the ontology is defined. These types of inter-annotator experiments that use the ontology as annotation guidelines are just now starting to be reported in the literature (Camon, Barrell, et al. 2004).

CONCLUSION:

The highly structured nature of ontologies and the semantics they represent will provide a valuable resource in natural language processing research in the foreseeable future. Text mining will be supported by enriching the text features of ontologies, improving indexing for search and retrieval and improving automatic mapping of objects to concepts. Life science ontologies themselves depend on the underlying text, since the biological concepts they represent are linked to the dynamic literature from which they are drawn. This would allow improved text mining to support efforts to automatically populate ontologies. The natural language processing community can also aid ontology design with experience in evaluation and inter-annotator studies to create semantic representations of greater utility to both human users and automatic systems.

REFERENCES

Morgan, A., L. Hirschman, et al. (2003). Gene Name Extraction Using FlyBase Resources. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan.

The FlyBase Consortium (1993). "FlyBase." GNome News(13): 19-20, http://www.geneontology.org.

The Gene Ontology Consortium (2000). "Gene Ontology: tool for the unification of biology." Nature Genetics(25): 25-29.

Valencia, A., C. Blaschke, et al. (2004). BioCreAtIvE Workshop Homepage. http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/.

P.W.Lord, R.D. Stevens, A. Brass, and C.A.Goble. Semantic Similarity Measures as Tools for Exploring the Gene Ontology. In 8th Pacific Symposium on Biocomputing (PSB), pages 601-612, 2003.

E.B. Camon, D.G. Barrell, E.C. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, R. Apweiler. “An evaluation of GO annotation retrieval for BioCreative and GOA”. Journal of Biomedical Informatics, *to be published*.