The Colorado OpenDMAP system: Building on Community Ontologies and a Community Platform for Biomedical Natural Language Processing

Karin Verspoor, William Baumgartner Jr., K. Bretonnel Cohen, Helen Johnson, and Lawrence Hunter
University of Colorado Denver, Center for Computational Pharmacology
karin.verspoor@ucdenver.edu, william.baumgartner@ucdenver.edu, kevin.cohen@gmail.com,
helen.linguist@gmail.com, larry.hunter@ucdenver.edu

Conceptual recognition is the process of mapping from natural language texts to a formal representation of the objects and predicates (together, the concepts) meant by the text. The history of attempts to build programs to do conceptual recognition dates back to at least 1967 [1]. Recent advances in the availability of high quality ontologies, in the ability to accurately recognize named entities in texts, and in language processing methods generally have made possible a significant advance in concept recognition, arguably the most difficult and general natural language processing task. We report on the design and implementation of OpenDMAP [2], an ontology-driven, integrated concept recognition system that advances the state of the art in biomedical information extraction by leveraging knowledge in ontological resources in the Open Biomedical Ontologies Foundry (OBO Foundry, http://obofoundry.org), integrating diverse text processing applications through the Apache Unstructured Information Management Architecture platform (UIMA; http://incubator.apache.org/uima), and using an expanded pattern language that allows mixing of syntactic and semantic elements and variable ordering.

OpenDMAP uses Protégé [3] to provide an object model for the possible concepts (predicates and objects) that might be found in a text. For example, the biological concept of *protein transport* is modeled as a class (called PROTEIN-TRANSPORT) and the relationship between a transport event and the protein transported in that event is represented as a slot in that class (called [TRANSPORTED-ENTITY]). Slots can take on values, which can be constrained to be instances of other classes. For example, the [TRANSPORTED-ENTITY] slot of the PROTEIN-TRANSPORT class is constrained to be an instance of either of the classes PROTEIN or MOLECULAR-COMPLEX. The semantics defined by the predicates and hierarchies in such ontologies provide a powerful tool for natural language processing. However, to our knowledge, OpenDMAP is the first system developed to exploit a community consensus ontology as the central organizing principle of an information extraction system. Other language processing systems have used either *ad hoc* conceptual representations developed for specific applications, or structured linguistic resources, such as WordNet [4], which do not meet the logical requirements for an ontology.

OpenDMAP utilizes a classic form of "semantic grammar," freely mixing text literals, semantically typed basal syntactic constituents, and semantically defined classes of entities. It is an extension of the Direct Memory Access Parsing (DMAP) paradigm described in [5] and [6]. The intimate connection between the ontology and the natural language processing system provides two significant advantages over prior information extraction systems generally. First, the output of the information extraction system is always constructed from elements of the ontology, ensuring that the knowledge representation is grounded with respect to a carefully constructed model of reality. Second, all of the knowledge used by the system to recognize concepts is structured by the ontology, such that semantic information about concepts inherently resolves some of the ambiguity issues faced by traditional lexicon-driven systems.

OpenDMAP has been applied to three biomedical information extraction problems and shown to be effective for each [2]: protein transport, protein-protein interaction and the expression of a gene in a particular cell type. It has been utilized for three shared task evaluations – BioCreAtIvE II [7], BioNLP09 [8] and Biocreative II.5 – and also as an expert in the Hanalyzer system [9]. OpenDMAP is freely available at http://bionlp.sourceforge.net/.

References

1. Sparck Jones K: Natural language processing: A historical review. Current Issues in Computational Linguistics: in Honour of Don Walker (Ed Zampolli, Calzolari and Palmer), Amsterdam: Kluwer 1994.
2. Hunter L, Lu Z, Firby J, Baumgartner WA Jr, Johnson HL, Ogren PV, Cohen KB: OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*. 2008 Jan 31; 9:78.
3. Noy NF, Crubezy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, Musen MA: Protege-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu Symp Proc 2004/01/20 edition. 2003, 953.
4. Fellbaum C: WordNet: An Electronic Lexical Database. MIT Press; 1998.
5. Martin C: Direct Memory Access Parsing. Yale University; 1992.
6. Fitzgerald W: Building Embedded Conceptual Parsers. Northwestern University; 1994.
7. Baumgartner WA, Lu Z, Johnson HL, Caporaso JG, Paquette J, Lindemann A, White EK, Medvedeva O, Cohen KB and Hunter L: Concept recognition for extracting protein interaction relations from biomedical text. Genome Biology 2008, 9(Suppl 2):S9.
8. Cohen KB, Verspoor K, Johnson HL, Roeder C, Ogren PV, Baumgartner WA, White EK, and Hunter L: Biological event extraction with a concept recognizer. Proceedings of the BioNLP09 workshop at HLT/NAACL; in press.
9. Leach SM, Tipney H, Feng W, Baumgartner WA Jr., Kasliwal P, et al.: Biomedical Discovery Acceleration, with Applications to Craniofacial Development. PLoS Comput Biol 5(3): e1000215; 2009.