A UIMA Wrapper for the NCBO Annotator

Christophe Roeder, Clement Jonquet, Nigam H. Shah, William A. Baumgartner Jr, Lawrence Hunter

Data and Text Mining:

A UIMA Wrapper for the NCBO Annotator

Christophe Roeder[1*], Nigam H. Shah[2], Clement Jonquet[2], William A Baumgartner Jr[1], and Lawrence Hunter[1]

[1]Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora, CO, USA

[2] Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine

Medical School Office Building, Room X-215 251 Campus Drive, Stanford, CA 94305-5479 USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: The UIMA framework and Web Services are emerging as useful tools for integrating biomedical text mining tools. This application note describes our work, which makes the NCBO Annotator available to UIMA workflows as a UIMA component.

Availability: This wrapper is freely available on the web at http://bionlp-uima.sourceforge.net/ as part of the center's UIMA tools distribution. It has been implemented in Java for support on Mac OS X, Linux and MS Windows

Contact: chris.roeder@ucdenver.edu

Integration and ease of installation are increasingly important concerns as the field of biomedical text mining tools grows in size and complexity. Issues include complex installation and integrating with other tools. Many tools are deployed as Web Services to avoid installation altogether. The NCBO's Annotator (Jonquet) is one such tool. It integrates many ontologies into an annotation service available on the web. Incremental users don't need to install it for themselves, just access the web service. UIMA (Ferruci) is an integration framework that makes combining disparate tools much easier. It provides a common user interface, common data representation and tool integration. The Center for Computational Pharmacology at the University of Colorado/SOM has adapted the NCBO Annotator to UIMA, making it available to UIMA projects.

The NCBO Annotator "automatically processes a piece of raw text to annotate (or tag) it with relevant ontology concepts and return the annotations. " (Jonquet) It makes use of much more than a single database or ontology and involves significant effort to integrate and maintain the data. Installing software and data locally would be cost prohibitive compared to remotely accessing an established instance. The annotator utilizes over 100 ontologies. They can be thought of as enriched term lists that include relationships and synonyms. One of the ontologies available is the Gene Ontology[1] and can be used to find references to cell components, biological processes, and molecular function. The annotator finds terms in submitted text that is related to the concepts in the ontologies and returns annotations describing them. For example, if "mitochondrion" appeared in the submitted text, the start and end character indexes of the word, the GO Onotology id, "39917", and an id for the concept in GO, "GO:0005739" would be returned. Such matches, direct matches, are found using the MGREP (Xuan) tool. The annotator makes use of the hierarchical nature of the ontologies as well as the UMLS[2] Thesaurus to provide more functionality. It can climb the ontology's hierarchy and report on more general concepts that relate to a particular word. "Intracellular membrane-enclosed organelle", "intracellular organelle", "organelle", and "cell component" are the higher members of the hierarchy starting with "mitochondrion". Such matches are called "is-a" matches. The UMLS also allows the annotator to navigate between ontologies and produce a broader range of results, including those from different forms of the word such as plurals ("mitochondria" would match "mitochondrion" for example) producing "mapping" matches.

This functionality is available over the web to users as a web service. A web service is similar to a website, but written for the use of computer programs. In this case, it is accessible to both humans and computers through an available web page[3]. For NLP projects that would make use of the functionality the annotator provides, the web service spares software developers the effort of procurement, installation, and maintenance of the code, data

*To whom correspondence should be addressed.

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

[2] http://umlsinfo.nlm.nih.gov/

[3] http://rest.bioontology.org/test_oba.html

and hardware involved. The ubiquity of the web and the simple protocols involved make it as easy to access a remote installation over the web to access a local installation.

The challenges for the bioinformatician extend beyond using a single service, to integrating with other tools to form a processing pipeline or workflow that accomplishes a greater goal like identifying and annotating protein-protein interaction events. These tools often work with different data formats for input and output, making the creation of a processing pipeline cumbersome. UIMA is a framework for integrating such tools into a common data format and interface. It provides a mechanism for running the tools in unison, and extensions provide for scaling to larger processing loads. Once a tool has been adapted to UIMA type system, it can be used in many different assemblies or pipelines with other tools also adapted to UIMA. The CCP has a collection (Baumgartner) of such adaptations or wrappers, which includes an adaptation of the NCBO Annotator to UIMA.

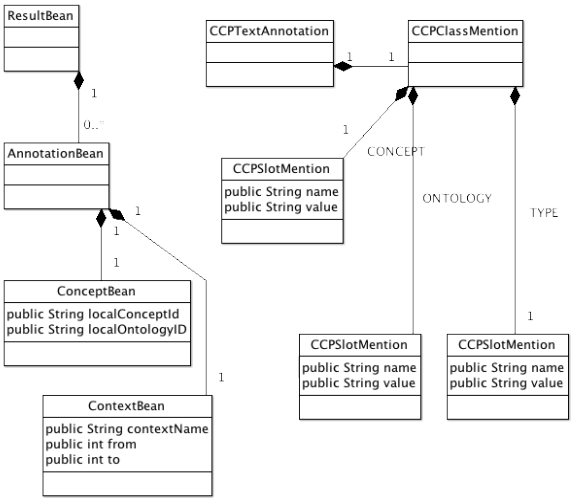

figure 1: UML sketches of NCBO (l) and CCP (r) representations

The UIMA wrapper for the annotator requests annotations in XML format. A partial UML[4] sketch of the structure of the returned XML is in the left of figure 1. Each match appears in an AnnotationBean and is linked to the document by the ContextBean. It shows the matching text and start and end (from, to) indexes of the span. The associated ConceptBean contains the IDs of the ontology and contained concept. The UIMA wrapper translates this to the structure sketched on the right of figure 1. A CCPTextAnnotation object contains the span information. An associated ClassMention has slots for each of the concept ID, the ontology ID and the match type. The match types are "DIRECT", for direct matches, "MAPPED" for matches created through the use of the UMLS thesaurus, and "IS-A" for more general results obtained by climbing the hierarchy of the ontologies. The "MAPPED" and "IS-A" types include more relevant detail documented elsewhere. (CITE?)

UIMA pipeline developers can include this component (with requisite sentence detector beforehand) in their pipelines. The component will query the NCBO Annotator for each sentence and add annotations for retrieved concepts. These concept annotations are then available for further processing by other components like OpenDMAP[5] for example, where they might be used in event recognition tasks.

The annotator and wrapper allow for control of the various match types through parameters in the web service. For example, climbing the isa hierarchy can be limited with a "levelMax" paramter. The ontologies involved searching the UMLS thesaurus can be limited with the "ontologiesToExpand" parameter as can the semantic Types used with "semanticTypes." "ontologiesToKeepInResult" restricts the results to a particular ontology. A full description is available at http://obs.bioontology.org/docs/oba/OBA_v1.1_documentation.htm.

ACKNOWLEDGEMENTS

Funding: This work was supported by National Institutes of Health grants [grant numbers R01GM083649, R01-008111, R01-009254 to LH]

REFERENCES

Ferrucci, D, *et al*. (2006) Towards an interoperability standard for text and multi-modal analytics. *IBM Res. Rep*., RC24122

Baumgartner WA, Jr, et al. *An open-source framework for large-scale, flexible evaluation of biomedical text mining systems*. J. Biomed. Discov. Collab. 2008;3:1.

Jonquet, M. A. Musen, N. H. Shah (2008) *Help will be provided for this task: Ontology-Based Annotator Web Service*. C. Technical Report

Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S.J., Meng, F.: Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In: BioLINK SIG: Linking Literature, Information and Knowledge for Biology, Vienna, Austria (Jul 2007) 55–58

[4] http://www.uml.org/

[5] http://opendmap.sourceforge.net/