A UIMA wrapper for the NCBO annotator

Christophe Roeder1,∗, Clement Jonquet2, Nigam H. Shah2, William A. Baumgartner Jr1, Karin Verspoor1 and Lawrence Hunter1

1Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora, CO and 2Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine, MSOB Room X-215, 251 Campus Drive, Stanford, CA 94305-5479, USA

ABSTRACT

Summary: The Unstructured Information Management Architecture (UIMA) framework and web services are emerging as useful tools for integrating biomedical text mining tools. This note describes our work, which wraps the National Center for Biomedical Ontology (NCBO) Annotator—an ontology-based annotation service—to make it available as a component in UIMA workflows.

Availability: This wrapper is freely available on the web at http://bionp-uima.sourceforge.net/ as part of the UIMA tools distribution from the Center for Computational Pharmacology (CCP) at the University of Colorado School of Medicine. It has been implemented in Java for support on Mac OS X, Linux and MS Windows.

Contact: chris.roeder@ucdenver.edu

Received on February 1, 2010; revised on April 12, 2010; accepted on May 9, 2010

Integration and ease of installation are increasingly important concerns as the number of biomedical text mining tools grows in size and complexity. Many tools are deployed as web services to avoid complex installation. The National Center for Biomedical Ontology (NCBO)’s Annotator (Jonquet, 2009) is one such tool. It integrates many biomedical ontologies into a concept annotation service available on the web. Instead of installing the tool locally, users outside of the NCBO can simply embed a web service client in their applications.

However, challenges for a bioinformatician extend beyond using a single service. Pipelines of tools are often assembled to accomplish a greater goal like identifying and annotating protein–protein interaction events. These tools work with different data formats for input and output, making the creation of a processing pipeline cumbersome. The Unstructured Information Management Architecture (UIMA) (Ferrucci, 2006) is a framework for integrating such tools into a common data representation and interface. It provides a mechanism for running the tools in unison and extensions for scaling to larger processing loads. UIMA is commonly used in the biomedical natural language processing (NLP) community, and recognition of ontology concepts is an important component of many text mining applications. Hence, adaptation of the NCBO Annotator to UIMA will likely result in broad adoption of the Annotator in biomedical text mining research.

Once a tool has been adapted to UIMA, it can be used in many different assemblies or pipelines with other tools also adapted to UIMA. The Center for Computational Pharmacology (CCP) at the University of Colorado School of Medicine has a collection of such adaptations or wrappers (Baumgartner, 2008), and has successfully used UIMA in the development of systems for participating in the BioCreative and BioNLP shared tasks. The CCP has now also 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, 2009). The Annotator accesses over 200 biomedical ontologies from the Unified Medical Language System (UMLS) Metathesaurus (http://www.nlm.nih.gov/research/umls/) and the NCBO BioPortal (http://bioportal.bioontology.org/). The biomedical ontologies used by the Annotator can be thought of as enriched term lists that include relationships and synonyms. One of the ontologies available is the Gene Ontology (GO), (http://www.geneontology.org/) which can be used to find references to cell components, biological processes and molecular functions. The Annotator identifies ontology terms in the submitted text and returns formally described annotations in the form of a Uniform Resource Identifier. For example, if mitochondrion appeared in the submitted text, the Annotator would return the start and end character indexes of the word, the GO ontology id, GO:0005739, and an id for the concept within GO, 39917, i.e., the GO id and an id for the concept within GO, 39917, and an id for the concept within GO, 39917, and an id for the concept within GO, 39917, and an id for the concept within GO, 39917. Such direct matches are found using the MGREP tool (Xuan, 2007). The Annotator enables the use of the hierarchical structure of the ontologies as well to provide more functionality: given a particular parameter setting, it can climb the ontology’s is-a hierarchy and return more general concepts that relate to a particular term. For instance, it could return the ancestors of mitochondrion, i.e., intracellular membrane-enclosed organelle, intracellular organelle, organelle and cell component. Such matches are called is-a matches. They are considered indirect matches because MGREP does not match mitochondrion with them, rather they are found through relationships in the ontology. The UMLS and BioPortal also allow the Annotator to search between ontologies using mappings to produce a broader range of results including those from different forms of the term such as plurals (mitochondrion) would match mitochondrion (for example).

The UIMA wrapper and the Annotator are actively being used in ongoing research that performs concept matching in text. The ultimate utility of this wrapper is to use the Annotator in more elaborate pipelines where the generated semantic annotations are used as input to downstream processing, for instance to help identify complex relationships between biological entities. For example, the CCP’s work in BioNLP’09 (Cohen, 2009) used similar methods to...
Annotator. A

They are unmarshalled from the returned XML by JAXB

This wrapper uses the CCP's type system, making it compatible with

MGREP. MAPPING is a match found in a second ontology based

for direct matches from the text to an ontology term found with

what kind of a match the Annotator used to find the concept.

attributes are ontology, concept and type. The type attribute tells

CCPSlotMention

objects that contain the attribute values. The

UIMA data structure, the Common Analysis Structure or CAS, using

using CCP wrappers. The topic of type system design is explored

system is used. U-Compare makes use of both a shared type system

can be re-used whenever another tool from the same foreign type

(Ferrucci, D. et al. 2008) and the Tsujii lab's U-Compare system (Kano, 2009),

involve some adaptation between type systems. Since no standard


parameter. A full description of parameters is available online (http://

the inheritance hierarchy can be limited with the levelMax parameter. A full description of parameters is available online (http://

parameters available in the web service. For example, climbing

such as a

and holds references to other objects describing the annotation

parameters available in the web service. For example, climbing

exception that stops the pipeline.

The wrapper is available as part of the CCP’s BioNLP-UIMA
distribution. It requires Java 1.6 and includes build scripts that run

The wrapper limits the size of the requests it makes on the web service by using sentence annotations created by an analysis engine upstream in the processing pipeline (we currently use the LingPipe toolkit for sentence analysis), and sending one sentence at a time. This keeps requests below the 300 word limit. The NCBO reports that with 10 simulated users, requests of 280 words take <5s. Our tests have been limited to five concurrent requests to leave some bandwidth for others. On the rare occasions the Annotator is unavailable, the wrapper waits for a short timeout and throws an exception that stops the pipeline.

The wrapper is available as part of the CCP's BioNLP-UIMA distribution. It requires Java 1.6 and includes build scripts that run on Macintosh and Linux variants. Windows support is forthcoming. Scripts download required third party jars and run the build. A sample pipeline is included ready to run after installation. Users outside our lab have installed the software quickly and easily. Further pipelines can be built using a UIMA-supplied GUI.

The NCBO Annotator provides annotations from a wealth of ontologies. Packaged as a web service, it is readily available to NLP researchers. With a UIMA adaptation provided by the CCP, it is now also available to the world of UIMA tools and pipelines.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers and Kevin Livingston for helpful discussion and review.

Funding: National Institutes of Health (5R01-GM083649-02, 2R01-LM008111-04A1, 2R01-LM009254-04 to L.H.); National Center for Biomedical Ontology (U54 HG004028).

Conflict of Interest: none declared.

REFERENCES


