MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-1191

Integrating UIMA Annotators in a Web-based Text Processing Framework

Xiang Chen[a], **Corey W. Arnold**[a]

[a] *Medical Imaging Informatics Group, University of California - Los Angeles, USA*

Abstract and Objective

The Unstructured Information Management Architecture (UIMA) [1] framework is a growing platform for natural language processing (NLP) applications. However, such applications may be difficult for non-technical users deploy. This project presents a web-based framework that wraps UIMA-based annotator systems into a graphical user interface for researchers and clinicians, and a web service for developers. An annotator that extracts data elements from lung cancer radiology reports is presented to illustrate the use of the system. Annotation results from the web system can be exported to multiple formats for users to utilize in other aspects of their research and workflow. This project demonstrates the benefits of a lay-user interface for complex NLP applications. Efforts such as this can lead to increased interest and support for NLP work in the clinical domain.

Keywords:

Natural language processing, Information extraction, Online text processing.

System Overview

We developed a web application framework for deploying UIMA annotators that provides a graphical user interface (GUI) that allows a user to select between different annotators and terminologies. After processing, users may either download annotations in XML format, or explore them in an interactive viewer, which displays the uploaded text along with options to select each of the annotations and view their corresponding features. The application is built using Spring model-view-controller (MVC), with the GUI designed to have all functions present on a single web page to minimize visual complexity. Asynchronous JavaScript and XML (AJAX) components allow annotation results to return to the client without requiring a page load. An instance of the UIMA annotation system is loaded on the server at all times, thus eliminating the need to repeatedly load large dependent libraries.

The framework features an application programming interface (API) that enables technical developers, who may not be familiar with UIMA-based annotators or may lack computing power, to integrate their applications with our system. The API side of the web application adheres to representational state transfer (REST) principles. In a stateless web environment, no annotation results are kept on the server once a piece of text has been processed. Instead, serialized UIMA Java objects are passed between server and client in order recall information from past annotation results. New annotators that follow UIMA standards may be uploaded to the web server as a JAR file with an associated XML Analysis Engine (AE) descriptor.

Results

To demonstrate the framework's utility, we have integrated two UMIA-based annotators for processing thoracic radiology reports to extract tumor characteristics in patients with lung cancer. First, the Mayo Clinic's Clinical Text Analysis and Knowledge Extraction System (cTAKES) [2] was imported into the application. Next, a less complex annotator consisting of a series of regular expressions was uploaded. Both annotators were then run via the interactive GUI and the API.

Discussion

Despite the success of the framework in modularizing UIMA annotators, several hurdles remain. The format in which a user wishes to export annotation results may vary widely. In the initial version of the system, we have focused on two simplified formats, XML and CSV. Future work includes allowing the user to more comprehensively select the features and annotations of interest to export, as well as integrating results with emerging annotation standards. Additionally, online tools will be developed to support interactive evaluation of annotator results.

UIMA-based annotators are important tools for researchers; however they require technical expertise to install and configure, posing a potential obstacle for individuals without the requisite experience. Making NLP results accessible to a larger number of researchers requires interfaces that cater to individuals with little text processing experience. Efforts such as this work may be used to illustrate the benefits of NLP, and lead to increased recognition and use by a wider audience. In addition, the platform provides a streamlined methodology for sharing annotators, enabling users to browse and test existing text processing solutions, rather than having to re-create them.

References

[1] Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering. 2004;**10**(3-4):327-48.

[2] Savova G, Masanz J, Ogren P, Zheng J, Sohn S, Kipper-Schuler K, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010;**17**(5):507.