BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2486–2487
doi:10.1093/bioinformatics/btq464

Data and text mining

Advance Access publication August 12, 2010

Text mining meets workflow: linking U-Compare with Taverna

Yoshinobu Kano[1,*], Paul Dobson[2], Mio Nakanishi[3], Jun'ichi Tsujii[1,4,5] and Sophia Ananiadou[4,5]

[1]Department of Computer Science, The University of Tokyo Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033 Japan, [2]School of Chemistry, The University of Manchester, Manchester M13 9PL, UK, [3]Department of Life Sciences (Biology), Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan, [4]School of Computer Science, The University of Manchester and [5]National Centre for Text Mining, 131 Princess St, M1 7DN, UK

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Text mining from the biomedical literature is of increasing importance, yet it is not easy for the bioinformatics community to create and run text mining workflows due to the lack of accessibility and interoperability of the text mining resources. The U-Compare system provides a wide range of bio text mining resources in a highly interoperable workflow environment where workflows can very easily be created, executed, evaluated and visualized without coding. We have linked U-Compare to Taverna, a generic workflow system, to expose text mining functionality to the bioinformatics community.

Availability: http://u-compare.org/taverna.html, http://u-compare.org

Contact: kano@is.s.u-tokyo.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 27, 2010; revised on August 5, 2010; accepted on August 8, 2010

1 INTRODUCTION

Owing to the large and rapidly growing body of literature in the biological sciences text mining approaches are increasingly important for the extraction and collation of data. The collation of text mining data with bioinformatics databases is a separate topic (Krallinger *et al*., 2008) which space precludes enlarging on here. Yet many text mining methods are difficult to integrate with other bioinformatics tools as development tends to focus upon mining performance more than implementing accessible interfaces. U-Compare (Kano *et al*., 2009) is an all-in-one text mining system based on the UIMA (Unstructured Information Management Architecture) framework (Ferrucci *et al*., 2006). U-Compare mainly consists of two parts: the world-largest ready-to-use UIMA component repository, and an all-in-one text mining platform.

In text mining data structures tend to be more complex (e.g. phrase structures) than in bioinformatics so interoperability, which guarantees data type compatibility as implemented in U-Compare, is required to create text mining workflows easily. With text mining workflows simplified, users would naturally seek integration of text mining workflows with other bioinformatics resources. Taverna (Hull *et al*., 2006) is a generic workflow construction system widely used in bioinformatics. We have developed mechanisms that allow users to embed any U-Compare workflow in a Taverna workflow. As U-Compare is designed to facilitate the construction of text mining workflows, it is far simpler to expose a complete U-Compare workflow to Taverna than to construct the equivalent workflow from individual components within Taverna, which lacks text mining specific data structures. It is therefore the easiest way to add text mining functionality to Taverna.

*To whom correspondence should be addressed.

2 SYSTEM FEATURES AND EXAMPLE WORKFLOWS

The U-Compare system itself is a stand-alone application. In this platform, users can create a workflow from the components in the repository, or any third-party UIMA components, in an easy drag-and-drop manner and compare, evaluate and visualize the workflow results. The entire system can be started by a single mouse click in the U-Compare website. Workflows can also be executed via the command line without the GUI based platform.

2.1 U-Compare Taverna plugin

The U-Compare Taverna plugin works with Taverna version 2.1.0. The user must specify two options: the U-Compare workflow to embed and a post-processing Beanshell script with proper I/O ports to appropriately reformat the output for further processes in Taverna. This post-processing script is executed as the final UIMA component in the U-Compare workflow. UIMA and U-Compare APIs can be used in the script. Upon running the Taverna workflow the U-Compare application starts and runs the specified text mining workflow automatically, then shows U-Compare GUIs such as statistics and visualizations of generated annotations. This plugin is implemented to download, install and update U-Compare automatically. Users are only required to install the plugin from the Taverna's menu by inputting the plugin URL. As running GUIs can be demanding when processing a large number of documents this plugin is mainly for testing and analyzing with results visualization.

Users can deploy two modes of workflow inputs to the U-Compare Taverna plugin. In the typical mode, the U-Compare workflow takes a collection reader as input, generates a list of annotated documents as output. We have also implemented another mode to link specifically with Taverna where the input is not the collection

reader but instead a list of String (depth 1) which is passed to the 'input_text' port of the U-Compare plugin.

2.2 U-Compare activity with command line mode

As any workflow can also be called via the command line we provide special UIMA components, which reads input text and writes generated annotations via the standard I/O streams. Using U-Compare's command line mode, we created an example Taverna workflow of a protein–protein interaction extraction, selected to show its usefulness in systems biology (Ananiadou *et al*., 2010). This workflow outputs a possible interaction network from the literature associated with a PubMed query. Extracted information is available in Supplementary Material. This example workflow is provided as a template for users to create their own workflow by imitating, reusing and modifying the interpretation part. Since U-Compare outputs results in a uniform format users only need to change the specific data types and their corresponding fields to create their own workflows, reusing most the template codes without modification. Figure 1 shows a diagram of the example Taverna workflow. A box labeled 'UCompare' corresponds to a Beanshell script activity, calls U-Compare via the command line. In this article, a mechanism to download, install and update the U-Compare system is also implemented, so there is no need to explicitly install anything. In parts prior to 'UCompare', this workflow retrieves documents from PubMed, passes them to the U-Compare activity, then interprets the results in parts following 'UCompare'.

An example run of the workflow, which sets 'PubMed_Query' and 'workflowPath' parameters in the Figure 1, and its result are given in Supplementary Material for the top 50 hits from the query 'saccharomyces AND "translation initiation" '. The 'workflowPath' is set to a U-Compare workflow given in the Supplementary Material, which runs 'UIMA Sentence Splitter' to detect sentence boundaries, 'ABNER' (Settles, 2005) to detect protein named entities and 'EventMine' (Miwa *et al*., 2010) to detect interaction events. Evaluations of the text mining tools are provided in U-Compare (Kano *et al*., 2009). This example run detects 677 non-unique entities, which by exact string matching correspond to 371 unique entity names, and 254 non-unique binding events.

2.3 U-Compare as a generic text mining workflow

The Taverna U-Compare workflow illustrated in Figure 1 can be used to run any U-Compare workflow, and does not require reworking for different text mining analyses. Whilst developing a new text mining workflow does involve configuration within the U-Compare environment, once developed, it can be deployed within the Taverna activity described here by resetting the 'workflowPath' value to the location of the new workflow's descriptor, relative to a directory [user home]/.UCompare/taverna/classpath-root/.

3 SUMMARY

U-Compare and Taverna focus upon different target domains, with the former utilizing a more strongly typed system specifically designed to handle the particular problems of text mining, while the latter offers a more generic solution for broader applications. By linking U-Compare with Taverna, we provide an easy way to create and embed complex text mining workflows within Taverna.

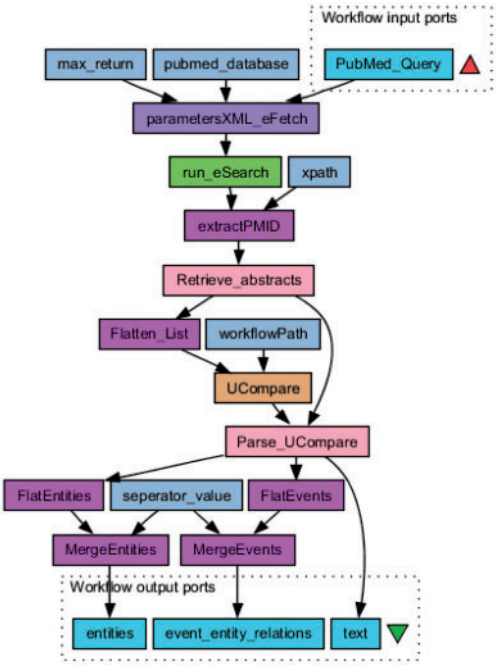

Fig. 1. An example Taverna workflow where U-compare activity is embedded, available in Supplementary Material and myExperiment (id 1377).

ACKNOWLEDGEMENTS

The authors thank Dr William Black at NaCTeM for his valuable advice, and the myGrid Taverna team for helping us to develop the Taverna plugin.

Funding: KAKENHI, Mext Japan (18002007, 21500130, in part); BBSRC (grant BB/F006012/1, in part).

Conflict of Interest: none declared.

REFERENCES

Ananiadou,S. *et al*. (2010) Event extraction for systems biology by text mining the literature. *Trends Biotechnol.*, **28**, 381–390.

Ferrucci,D. *et al*. (2006) Towards an interoperability standard for text and multi-modal analytics. *IBM Research Report*, **RC24122**, W0611–188.

Hull,D. *et al*. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.

Kano,Y. *et al*. (2009) U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, **25**, 1997–1998.

Krallinger,M. *et al*. (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9** (Suppl. 2), S8.

Miwa,M. *et al*. (2010) Event extraction with complex event classification using rich features. *J. Bioinform. Comput. Biol.*, **8**, 131–146.

Settles,B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, **21**, 3191–3192.