BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 15 2009, pages 1997–1998
doi:10.1093/bioinformatics/btp289

Data and text mining

U-Compare: share and compare text mining tools with UIMA

Yoshinobu Kano[1,*], William A. Baumgartner Jr[2], Luke McCrohon[1], Sophia Ananiadou[3,4], K. Bretonnel Cohen[2], Lawrence Hunter[2] and Jun'ichi Tsujii[1,3,4]

[1]Department of Computer Science, University of Tokyo Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033 Japan,
[2]Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora, CO, USA,
[3]School of Computer Science, University of Manchester and [4]National Centre for Text Mining, 131 Princess St.,
M1 7DN, UK

Received on February 7, 2009; revised on March 26, 2009; accepted on April 24, 2009
Advance Access publication May 4, 2009
Associate Editor: Jonathan Wren

ABSTRACT

Summary: Due to the increasing number of text mining resources (tools and corpora) available to biologists, interoperability issues between these resources are becoming significant obstacles to using them effectively. UIMA, the Unstructured Information Management Architecture, is an open framework designed to aid in the construction of more interoperable tools. U-Compare is built on top of the UIMA framework, and provides both a concrete framework for out-of-the-box text mining and a sophisticated evaluation platform allowing users to run specific tools on any target text, generating both detailed statistics and instance-based visualizations of outputs. U-Compare is a joint project, providing the world's largest, and still growing, collection of UIMA-compatible resources. These resources, originally developed by different groups for a variety of domains, include many famous tools and corpora. U-Compare can be launched straight from the web, without needing to be manually installed. All U-Compare components are provided ready-to-use and can be combined easily via a drag-and-drop interface without any programming. External UIMA components can also simply be mixed with U-Compare components, without distinguishing between locally and remotely deployed resources.

Availability: http://u-compare.org/

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

1 INTRODUCTION

In the biomedical domain, an increasing number of text mining tools have been developed, and some of these are now ready for biologists and database curators to use for their own needs (Ananiadou *et al.*, 2006). However, it is still very difficult to integrate independently developed tools into an aggregate application. Difficulties are caused not only by differences in programming platforms or different input/output formats, but also by the lack of higher-level interoperability among modules developed by different groups.

UIMA (Ferrucci *et al.*, 2006), Unstructured Information Management Architecture, is a robust and flexible framework that facilitates interoperability between tools. UIMA is currently an Apache open source project with its specification in OASIS, being widely used. Although UIMA provides rich functionality, it is intended to be a general framework. To apply the framework to a specific domain, e.g. text mining, users need to create a UIMA-compatible type system, which defines the data types used by their tools. U-Compare provides such a type system,which aims to allow for comparison of resources, and conversion between different users' individual type systems (Kano *et al.*, 2008). The U-Compare type system covers a wide range of text mining concepts to help bridge existing type systems. Although U-Compare uses its own type system for included components, it is still compatible with UIMA generally and users can choose to use their own type systems if they so desire.

U-Compare is built on top of UIMA, and is a joint project between the University of Tokyo, the University of Colorado School of Medicine and the National Centre for Text Mining at the University of Manchester. We have developed the world's largest set of type system compatible UIMA components (Table 1), which are integrated under the U-Compare type system. To demonstrate the power of these included components, U-Compare includes a number of predefined workflows demonstrating the range of possible combinations. We are continuously working to increase the number of included components and are collaborating with other research groups to achieve this. Other significant projects that aim to provide collections of UIMA components include those at: the CMU UIMA Component Repository (http://uima.lti.cs.cmu.edu/), JCoRe (Hahn *et al.*, 2008) and the BioNLP UIMA Repository (Baumgartner *et al.*, 2008).

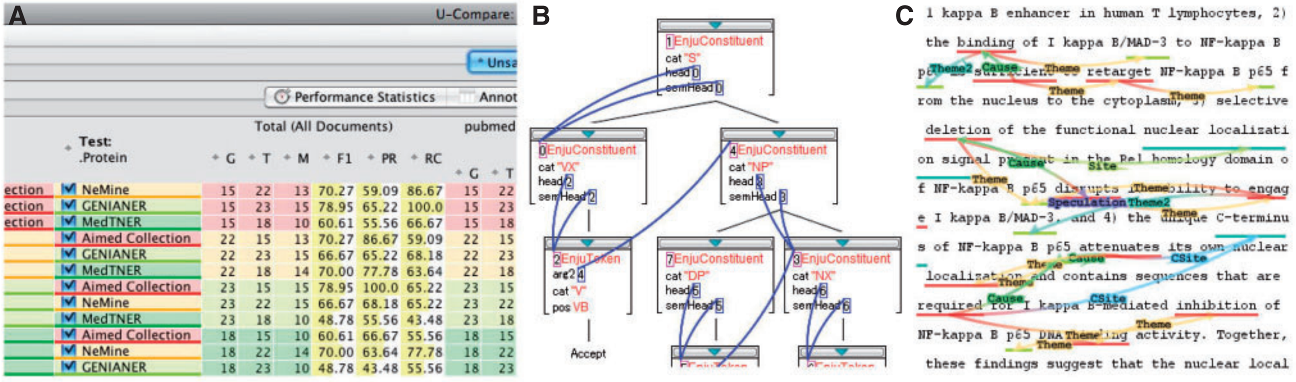
2 SYSTEM FEATURES

The entire U-Compare system can be launched by a single click from the U-Compare website, without any explicit installation operation, under any Java-enabled OS. The U-Compare system and components are individually downloaded on demand, cached and updated automatically via the Java Web Start technology. Within U-Compare any UIMA component can be used by drag-and-dropping it from the U-Compare component library into the appropriate location in the workflow manager. This works regardless of whether the component is a locally or remotely deployed service.

2.1 Workflow management and combinatorial comparison

UIMA workflows are built from UIMA components, which can be nested as parts of an *aggregate component*. The U-Compare

[*]To whom correspondence should be addressed.



	Test: .Protein	Total (All Documents) G	T	M	F1	PR	RC	pubmed G	T
ection	☑ NeMine	15	22	13	70.27	59.09	86.67	15	22
ection	☑ GENIANER	15	23	15	78.95	65.22	100.0	15	23
ection	☑ MedTNER	15	18	10	60.61	55.56	66.67	15	18
	☑ Aimed Collection	22	15	13	70.27	86.67	59.09	22	15
	☑ GENIANER	22	23	15	66.67	65.22	68.18	22	23
	☑ MedTNER	22	18	14	70.00	77.78	63.64	22	18
	☑ Aimed Collection	23	15	15	78.95	100.0	65.22	23	15
	☑ NeMine	23	22	15	66.67	68.18	65.22	23	22
	☑ MedTNER	23	18	10	48.78	55.56	43.48	23	18
	☑ Aimed Collection	18	15	10	60.61	66.67	55.56	18	15
	☑ NeMine	18	22	14	70.00	63.64	77.78	18	22
	☑ GENIANER	18	23	10	48.78	43.48	55.56	18	23

I kappa B enhancer in human T lymphocytes, 2)
the binding of I kappa B/MAD-3 to NF-kappa B
p65 is sufficient to retarget NF-kappa B p65 f
rom the nucleus to the cytoplasm, 3) selective
deletion of the functional nuclear localizati
on signal present in the Rel homology domain o
f NF-kappa B p65 disrupts its ability to engag
e I kappa B/MAD-3, and 4) the unique C-terminu
s of NF-kappa B p65 attenuates its own nuclear
localization and contains sequences that are
required for I kappa B-mediated inhibition of
NF-kappa B p65 DNA binding activity. Together,
these findings suggest that the nuclear local

Fig. 1. Screenshots of (**A**) U-Compare Statistics Viewer showing comparison between AImed corpus and three NERs; (**B**) U-Compare Tree and Feature Structure Visualizer showing an HPSG syntactic tree; and (**C**) U-Compare Graphical Annotation Viewer showing biological event annotations.

Table 1. Partial list of currently ready-to-use components in U-Compare

Component type	Component names
Collection readers	AImed, Bio1, BioIE, Texas, Yapex, NLPBA
Sentence detectors	Genia, LingPipe, NaCTeM, OpenNLP, UIMA
Tokenizers	GENIA, OpenNLP, UIMA, PennBio
POS taggers	GENIA, LingPipe, OpenNLP, Stepp
Syntactic parsers	Enju HPSG Parser, OpenNLP Parser, Stanford Parser
Relation extracters	Akane++, BioNLP '09 Shared Task Format Reader
Named entity recognizers	ABNER, GENIA Tagger, NeMine, MedTNER, MedTNER-M, LingPipe Entity Tagger, OpenNLP

workflow manager allows users to create these workflows via an easy drag-and-drop interface. Workflows or configured components can be saved, reused and transferred between users.

U-Compare provides a special *parallel flow* component, which can be used to make *comparison workflows* to compare the outputs of tool and corpus combinations. U-Compare automatically decides possible combinations of comparison components (Kano *et al.*, 2008), based on a given workflow and component input/output types.

2.2 Evaluation, statistics and visualization

Running workflows requires just a single click. When the workflow completes, annotation instance counts and runtime performance statistics are automatically displayed. For comparison workflows, comparison statistics such as F1, precision and recall scores are also given (Figure 1A).

Additionally, visualizations of annotations are also made available (Figure 1C). For more complex visualizations of syntactic trees and feature structures (Figure 1B), special visualization components are included.

2.3 Developer APIs

In addition to the UIMA official Java/C++ APIs, U-Compare provides a simpler interface which allows developers access to a UIMA workflow via the standard I/O streams or via stored files. Workflows can be launched directly from the command line, taking any required inputs from *standard in* and outputting directly to *standard out*. Special components can also be embedded into workflows that communicate with developer's native tools via the simplified U-Compare interface.

3 SUMMARY AND FUTURE DIRECTIONS

U-Compare currently provides the world's largest collection of type-system-compatible UIMA resources. Combining and comparing these resources is made simple by a user-friendly drag-and-drop interface, a fully interoperable type system and a range of graphical tools for analyzing their outputs. In the future we will continue collecting components and enhancing our type system, through collaboration with other research groups, and hope to include tools such as relation extractors, more syntactic parsers, collection readers for corpora in which biological relations annotated, machine learning tools and more typical biological workflows.

Funding: This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan); and the National Institutes of Health [R01LM009254, R01LM008111, and R01GM083649]. The UK National Centre for Text Mining is funded by JISC.

Conflict of Interest: none declared.

REFERENCES

Ananiadou,S. *et al*. (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol*., **24**, 571–579.

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

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

Hahn,U. *et al*. (2008) An overview of JCoRe, the JULIE lab UIMA component repository, In *Proceedings of LREC'08 Workshop, Towards Enhanced Interoperability for Large HLT Systems*: UIMA for NLP, 1–8.

Kano,Y. *et al*. (2008) Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example. *Proc. Pac. Symp. Biocomput*., **13**, 616–627.