

# U-Compare Platform

## 1. BASIC INFORMATION

### *Tool name*

U-Compare platform

### *Overview and purpose of the tool*

The purpose of the U-Compare platform (Kano et al., 2011; Kano et al; 2009) is to facilitate easy and rapid development and evaluation of NLP and text mining systems. It includes utilities (including a graphical user interface, the U-Compare workbench, see separate record in META-SHARE) to create workflows from individual, interoperable NLP tools and resources, a customisable system to create and evaluate different workflows, and different utilities to visualise different types of annotations produced by workflows. U-Compare is packaged with the world's largest repository of UIMA components. This repository, which originally consisted largely of tools for processing English biomedical text, in being considerably enlarged as ongoing work, to include tools that can operate on a number of European languages and multilingual tools (Ananiadou et al, 2011; Thompson et al, 2011).

### *A short description of the algorithm*

U-Compare is a platform rather than a tool that performs a specific purpose. Hence, there is no specific algorithm to describe. However, this section provides some further implementation details.

U-Compare is built on top of the Unstructured Management Architecture (UIMA)<sup>1</sup> (Ferruci et al., 2006), which is a generic framework widely adopted by the NLP community to improve the interoperability of tools/components. U-Compare provides several tools that are necessary for NLP application development, but which are not provided by UIMA, due to the more general nature of UIMA.

U-Compare requires that tools and resources used in workflows are UIMA components. UIMA components are interoperable, in the sense that their input/output mechanisms are standardised – they must obtain their input by reading annotations from a data structure, the Common Analysis Structure (CAS), that is accessible to all components in a workflow, and output consists of adding new annotations to the CAS, or updating existing ones. Existing tools can be “wrapped” as UIMA components by writing code to convert their input/output formats to those required by UIMA.

The U-Compare platform allows workflows to be created simply by specification of which components to run, and in which order. This is helped by the Workbench

---

<sup>1</sup> <http://uima.apache.org>

graphical user interface (see separate META-SHARE record), which allows users to create workflows using drag-and-drop actions. Comparison workflows can also be created to compare and evaluate the outputs of several different workflows that perform the same task, possibly against a gold standard annotated corpus if this is available, with results displayed in graphical form.

Whilst U-Compare workflows can be created using any UIMA components, U-Compare also defines its own type system (described in a separate META-SHARE record), which aims to facilitate semantic interoperability between components. By “types”, we mean the categories of annotations that are input/output by the UIMA components. UIMA itself does not define or impose the use of a particular type system. This means that interoperability between components produced by different developers can be difficult to achieve, as they may use different sets of types as input/output. The U-Compare type system aims to resolve this problem, by providing a hierarchical *sharable* system of the most common types of annotations produced by NLP tools. The idea is that compatibility between a large number of UIMA components can be achieved, at least at an intermediate level of the hierarchy, through mapping of types to this type system. All components in the U-Compare repository are compatible with the U-Compare type system.

The U-Compare platform can be used independently of the U-Compare Workbench Graphical User Interface to run workflows directly from the command line. Additionally, U-Compare provides UIMA components for standard I/O streams that communicate in a simple standoff annotation format, which allows easy embedding of workflows into other systems, regardless of programming language.

## 2. TECHNICAL INFORMATION

### ***Software dependencies and system requirements***

The U-Compare platform can be used in any environment in which Java 6 is available. At least the first time the system is run, an Internet connection is required, since the most up-to-date relevant files are downloaded from the internet.

### ***Installation***

No specific installation is required. U-Compare can be started directly from the Internet by clicking on the “Start U-Compare” button on this page:

<http://www.nactem.ac.uk/ucompare/index.html>

However, it is preferable to start U-Compare from the command line, by downloading the file `UCLoader.class` from <http://u-compare.org/downloads/UCLoader.class>.

See also <http://www.nactem.ac.uk/ucompare/launch.html> for more information

### **Execution instructions**

From the command line, the U-Compare workbench is started by running the `UCLoader.class` file, e.g.

```
java -Xms700m -Xmx1000m UCLoader
```

The `-Xms` and `-Xmx` specified the minimum and maximum memory allocated to U-Compare. The more memory is allocated, the quicker U-Compare will run. Note that the first time U-Compare is launched, relevant files will be downloaded from the internet. Therefore, the first time the system is launched, it may take a considerable amount of time to start up.

The default behaviour of UCLoader is to start the U-Compare Workbench interface. However, other options can be specified that allow workflows to be run from the command line, without starting up the interface. The general way to run a workflow from the command line is as follows:

```
java -cp . -XmsXXXm -XmxXXXm  
-Djavaws.workflow.path="path/to/yourworkflow.xml" UCLoader --jnlp  
http://u-compare.org/lib/u-compare-runworkflow.jnlp
```

“`-cp`” is the Java VM option to specify your classpath (in this case the current directory “`.`” is specified). You should include `UCLoader.class` in your classpath.

“`-Xms`” and “`-Xmx`” are the Java VM options to specify the amount of heap memory allocation.

An example workflow that can be run on the command line, together with more details, are provided here:

[http://www.nactem.ac.uk/ucompare/developerguide/Command\\_Line\\_Mode\\_without\\_U.html](http://www.nactem.ac.uk/ucompare/developerguide/Command_Line_Mode_without_U.html)

### *Input/Output data formats*

#### **Input data formats**

Workflows that are run in the platform obtain input data via Collection Reader components. Currently, only text may be read in, but in the future, files of other modalities, e.g. speech, are planned. A collection reader reads a text or set of texts, which may be unannotated or may already contain annotations. The U-Compare library includes several types of collection reader, e.g. to read in text from an input window, from a directory of files or, in the case of a workflow being run from the command line, from standard input. Several corpus-specific readers are also provided, that read in annotated data.

#### **Output data format**

The result of running a workflow in U-Compare is a set of annotations added to the UIMA CAS. The contents of the CAS may be exported to different formats, e.g., to

files in the XMI (XML Metadata Interchange) format (Grose et al, 2002) or inline XML format, or to standard output, in the case that the workflow is run from the command line.

The U-Compare platform also includes annotation visualisation tools, which display annotations produced by the workflow as underlines and arcs superimposed on the document text, and highlight differences discovered during component comparison, as well statistics relating to the comparison, such as F-score, precision, recall, etc.

### Integration with external tools

The platform does not require the use of any external tools. As mentioned above, workflows created using the U-Compare platform can be embedded into other applications.

## 3. CONTENT INFORMATION

Figure 1 illustrates annotated visualisation in U-Compare, while Figure 2 illustrates the display of workflow comparison/evaluation results. More details on how to create and evaluate workflows in the U-Compare Workbench interface are provided in the narrative documentation that accompanies the U-Compare workbench META-SHARE record.

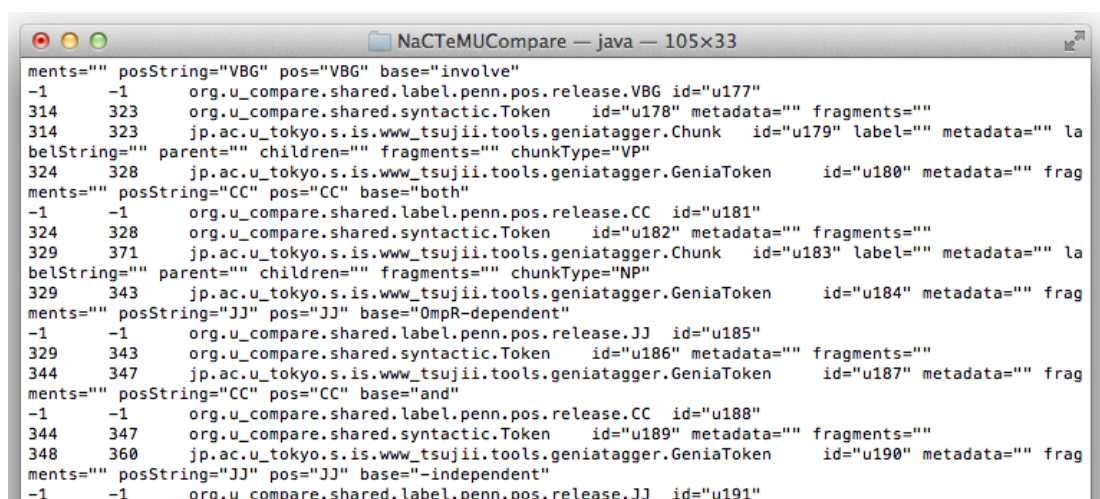
1 kappa B enhancer in human T lymphocytes, 2) the binding of I kappa B/MAD-3 to NF-kappa B p65. 3) I kappa B/MAD-3 retarget NF-kappa B p65 from the nucleus to the cytoplasm, 4) selective deletion of the functional nuclear localization signal present in the Rel homology domain of NF-kappa B p65 disrupts its ability to engage I kappa B/MAD-3, and 4) the unique C-terminus 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

Figure 1: Annotation visualisation

Assumed Gold Standard	Comparison Components	Total (All Documents)					
		Boundary Match					
▼ .Protein	▲ .Protein	↕ G	↕ T	↕ M	↕ F1	↕ PR	↕ RC
<input checked="" type="checkbox"/> Aimed	<input checked="" type="checkbox"/> ABNER-NLPBA	15	23	15	78.95	65.22	100.0
<input checked="" type="checkbox"/> ABNER-NLPBA	<input checked="" type="checkbox"/> Aimed	23	15	15	78.95	100.0	65.22
<input checked="" type="checkbox"/> Aimed	<input checked="" type="checkbox"/> ABNER-BioCreative	15	21	15	83.33	71.43	100.0
<input checked="" type="checkbox"/> ABNER-BioCreative	<input checked="" type="checkbox"/> Aimed	21	15	15	83.33	100.0	71.43

Figure 2: Workflow comparison/evaluation

In Figure 2, text based output is shown, as a result of running a workflow from command line. For each annotation, the annotation offsets, annotation type and annotation attributes are displayed.



```
ments="" posString="VBG" pos="VBG" base="involve"
-1 -1 org.u_compare.shared.label.penn.pos.release.VBG id="u177"
314 323 org.u_compare.shared.syntactic.Token id="u178" metadata="" fragments=""
314 323 jp.ac.u_tokyo.s.is.www_tsujii.tools.geniatagger.Chunk id="u179" label="" metadata="" la
belString="" parent="" children="" fragments="" chunkType="VP"
324 328 jp.ac.u_tokyo.s.is.www_tsujii.tools.geniatagger.GeniaToken id="u180" metadata="" frag
ments="" posString="CC" pos="CC" base="both"
-1 -1 org.u_compare.shared.label.penn.pos.release.CC id="u181"
324 328 org.u_compare.shared.syntactic.Token id="u182" metadata="" fragments=""
329 371 jp.ac.u_tokyo.s.is.www_tsujii.tools.geniatagger.Chunk id="u183" label="" metadata="" la
belString="" parent="" children="" fragments="" chunkType="NP"
329 343 jp.ac.u_tokyo.s.is.www_tsujii.tools.geniatagger.GeniaToken id="u184" metadata="" frag
ments="" posString="JJ" pos="JJ" base="OmpR-dependent"
-1 -1 org.u_compare.shared.label.penn.pos.release.JJ id="u185"
329 343 org.u_compare.shared.syntactic.Token id="u186" metadata="" fragments=""
344 347 jp.ac.u_tokyo.s.is.www_tsujii.tools.geniatagger.GeniaToken id="u187" metadata="" frag
ments="" posString="CC" pos="CC" base="and"
-1 -1 org.u_compare.shared.label.penn.pos.release.CC id="u188"
344 347 org.u_compare.shared.syntactic.Token id="u189" metadata="" fragments=""
348 360 jp.ac.u_tokyo.s.is.www_tsujii.tools.geniatagger.GeniaToken id="u190" metadata="" frag
ments="" posString="JJ" pos="JJ" base="-independent"
-1 -1 org.u_compare.shared.label.penn.pos.release.JJ id="u191"
```

Figure 2: Ouput of a workflow in the command line

## 4. LICENCE

The U-Compare platform is released under a dual license, the LGPL open source license (<http://www.gnu.org/licenses/lgpl.html>) or a commercial license. Please use the contact details below if you are interested in obtaining a commercial licence.

## 5. ADMINISTRATIVE INFORMATION

### Contact

For further information, please contact Sophia Ananiadou:

[sophia.ananiadou@manchester.ac.uk](mailto:sophia.ananiadou@manchester.ac.uk)

## 6. REFERENCES

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T. K., Day, P. J. R., Keane, J., Jackson, D. and Pettifer, S.. (2011). "Towards Interoperability of European Language Resources". *Ariadne*, 67

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W., Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Groose, T.J. Doney, G.C. and Brodsky, S.A. (2002). *Mastering XMI. Java Programming with XMI, XML, and UML*. John Wiley & Sons, Inc.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinformatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.

Thompson, P., Kano, Y., McNaught, J., Pettifer, S., Attwood, T. K., Keane, J. and Ananiadou, S. (2011). Promoting Interoperability of Resources in META-SHARE. In *Proceedings of the IJCNLP Workshop on Language Resources, Technology and Services in the Sharing Paradigm (LRTS)*, pp. 50-58