UIMA/U-Compare OpenNLP Tokenizer

1. BASIC INFORMATION

Tool name

U-Compare OpenNLP Tokenizer

Overview and purpose of the tool

This is a UIMA[1] (Ferrucci et al., 2006) wrapper for the OpenNLP Tokenizer tool. It splits English sentences into individual tokens.

The tool forms part of the in-built library of components provided with the U-Compare platform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record)[2] for building and evaluating text mining workflows. The U-Compare Workbench (see separate META-SHARE record), which provides a graphical drag-and drop interface for the rapid creation of workflows.

A short description of the algorithm

OpenNLP tools[3] are trained using machine-learning methods. The tool provided uses the pre-trained model for English, available on the OpenNLP SourceForge website: http://opennlp.sourceforge.net/models-1.5/

2. TECHNICAL INFORMATION

Software dependencies and system requirements

In order to run U-Compare, Java 6 must be installed.

The UIMA component calls a web service. Hence, internet access is required.

Installation

There is no specific installation for U-Compare. The file UCLoader.class should be downloaded from http://u-compare.org/downloads/UCLoader.class

Execution instructions

U-Compare is started by running UCLoader.class from the command line. Since U-Compare can consume a large amount of memory, it is suggested to specify minimum and maximum memory usage when running U-Compare, as in the following example:

[1] http://uima.apache.org/

[2] http://nactem.ac.uk/ucompare/

[3] http://opennlp.apache.org/

```
java -jar  -Xms700m -Xmx 1000m UCLoader
```

The memory usage can be adjusted, but note that a minimum memory usage of 256 MB is recommended. Please also note that when U-compare is first started for the first, a large number of files will be downloaded, and so it will take some time to start. Subsequent launches will be quicker.

Once U-Compare has been started, the sentence detector tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record “U-Compare Workbench” for more details

Input/Output data formats

Input data formats

The tool requires sentence split text as input. Thus, the UIMA Common Analysis Structure (CAS) must contain sentence annotations before this component is run. In a UIMA workflow, this could be achieved either by executing a component that performs sentence splitting prior to this component, or otherwise reading in a corpus of documents that already contains sentence annotations.

Output data format

Since the purpose of the tool is to detect tokens in the text, the result of running the tool is that an annotation corresponding to each token in the text is thus added to the CAS. Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

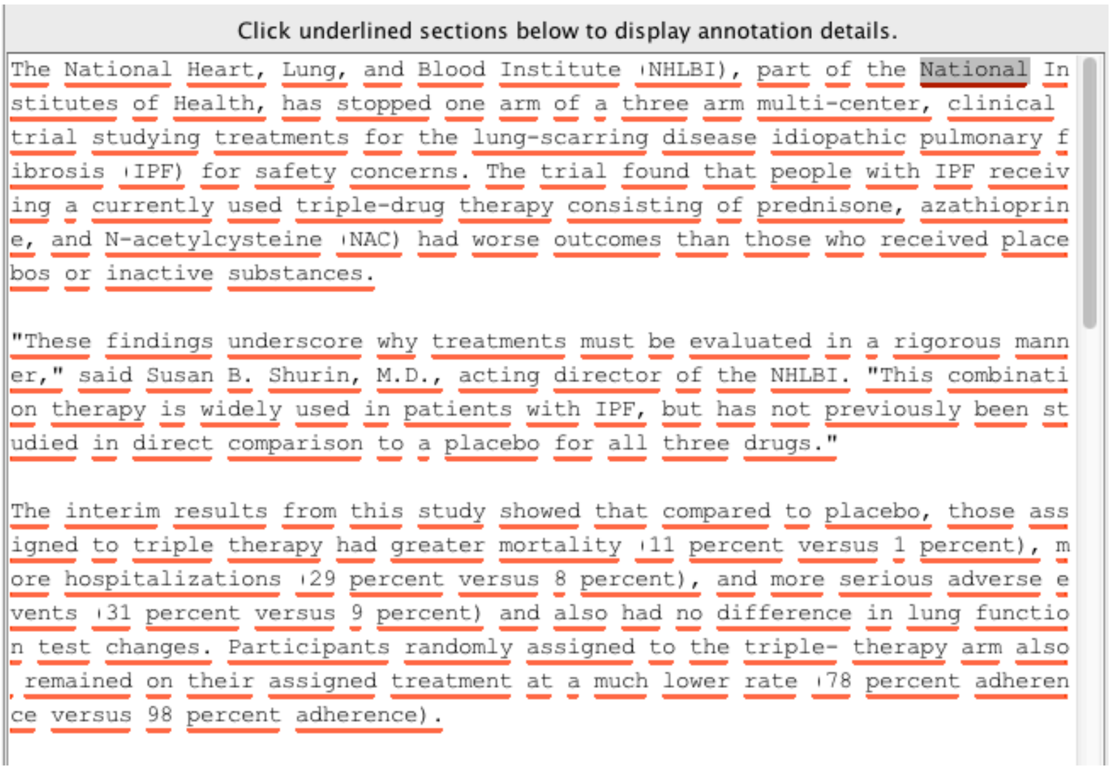
Integration with external tools

The tool can be run as part of a UIMA workflow, either using U-Compare or otherwise. For instructions of how to include components in UIMA workflows outside of U-Compare, see:

http://nactem.ac.uk/ucompare/developerguide/Using_U_Compare_Components_.html

3. CONTENT INFORMATION

Figure 1 shows the output of the tool in the U-Compare workbench. Each token recognised is separately underlined. The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)



Click underlined sections below to display annotation details.

```
The National Heart, Lung, and Blood Institute (NHLBI), part of the National In
stitutes of Health, has stopped one arm of a three arm multi-center, clinical
trial studying treatments for the lung-scarring disease idiopathic pulmonary f
ibrosis (IPF) for safety concerns. The trial found that people with IPF receiv
ing a currently used triple-drug therapy consisting of prednisone, azathioprin
e, and N-acetylcysteine (NAC) had worse outcomes than those who received place
bos or inactive substances.

"These findings underscore why treatments must be evaluated in a rigorous mann
er," said Susan B. Shurin, M.D., acting director of the NHLBI. "This combinati
on therapy is widely used in patients with IPF, but has not previously been st
udied in direct comparison to a placebo for all three drugs."

The interim results from this study showed that compared to placebo, those ass
igned to triple therapy had greater mortality (11 percent versus 1 percent), m
ore hospitalizations (29 percent versus 8 percent), and more serious adverse e
vents (31 percent versus 9 percent) and also had no difference in lung functio
n test changes. Participants randomly assigned to the triple- therapy arm also
 remained on their assigned treatment at a much lower rate (78 percent adheren
ce versus 98 percent adherence).
```

Figure 1: Output of the U-Compare OpenNLP Tokenizer in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 2.7 seconds.

4. LICENCES

The UIMA wrapper code, the underlying OpenNLP Tokenizer tool and the UIMA framework are all licensed using the Apache licence. See “Apache-licence.txt” in the “Licences” directory within the distribution.

5. ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

Copyright statement and information on IPR

The OpenNLP Sentence Detector must be used in compliance with the Apache Licence: http://www.apache.org/licenses/

6. REFERENCES

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.

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. *Bioinfomatics*, 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.