UIMA/U-Compare GENIA Sentence Splitter

1. BASIC INFORMATION

Tool name

UIMA/U-Compare GENIA Sentence Splitter

Overview and purpose of the tool

The purpose of the tool is to detect sentence boundaries in English text. It is trained on the GENIA corpus of biomedical abstracts (Kim et al., 2003) and so is particularly suitable for splitting sentences in biomedical texts.

The tool is provided as a UIMA[1] (Ferrucci et al., 2006) component, which 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

This sentence detector is trained on the GENIA corpus (Kim et al., 2003), using machine learning methods.

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:

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

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

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

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 operates on plain, unannotated text. Thus, the UIMA Common Analysis Structure (CAS) should contain the text to be analysed prior to the tool being executed. In a UIMA workflow, this could be achieved by reading in a single text or corpus of text. For example, U-Compare provides collection readers that can read in text from an input box, or otherwise read a directory of texts.

Output data format

The tool detects the boundaries of sentences and, for each sentence, adds an annotation to the UIMA CAS corresponding to the sentence. 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.

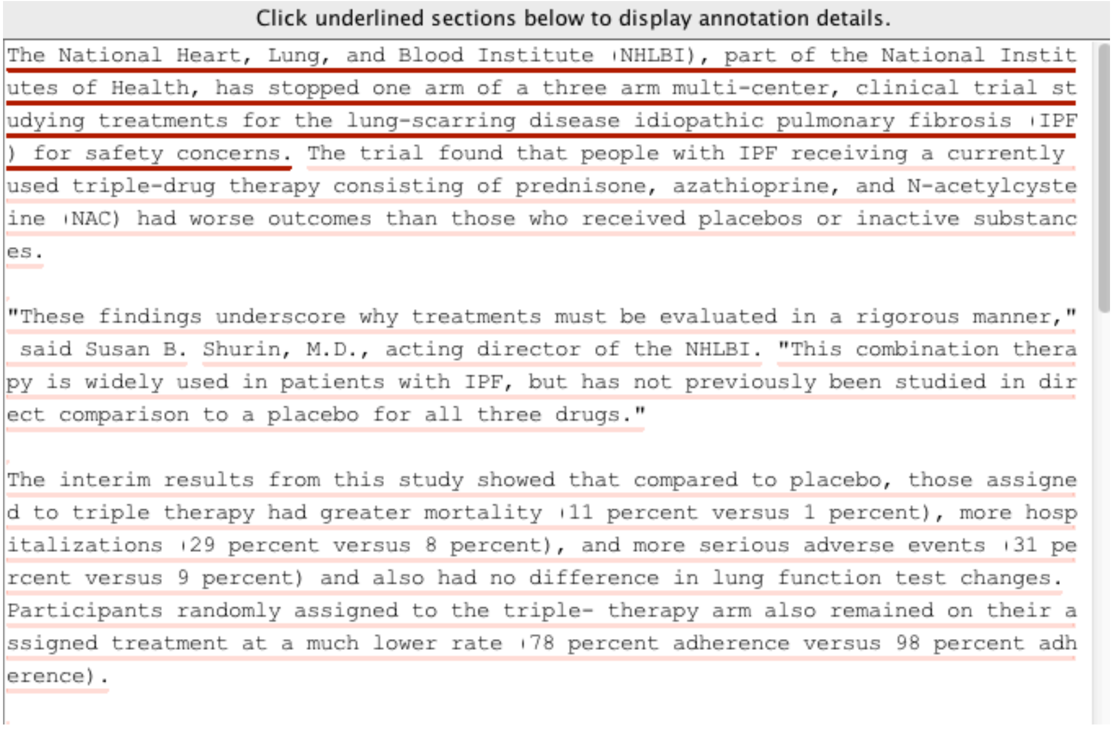
Integration with external tools

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. One of the sentences is highlighted. 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 Institutes of Health, has stopped one arm of a three arm multi-center, clinical trial studying treatments for the lung-scarring disease idiopathic pulmonary fibrosis (IPF) for safety concerns. The trial found that people with IPF receiving a currently used triple-drug therapy consisting of prednisone, azathioprine, and N-acetylcysteine (NAC) had worse outcomes than those who received placebos or inactive substances.

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

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

Figure 1: Output of the U-Compare GENIA Sentence Detector in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 2.43 seconds.

4. LICENCES

a) The UIMA wrapper code is licensed using the NaCTeM Software Licence Agreement (standard non-commercial use) – see “GENIA-Sentence-Splitter-U-Compare-licence.pdf” in the “licences” directory. Please contact us using the details below if you require a commercial licence.

b) The underlying GENIA Tagger web service is licensed using the NaCTeM Web Service Licence Agreement (standard non-commercial use)– see “GENIA-Sentence-Splitter-licence.pdf” in the “licences” directory. Please contact us using the details below if you require a commercial licence.

c) The UIMA framework is licenced using the Apache licence. Please see “Apache.txt” in the licenses directory.

5. ADMINISTRATIVE INFORMATION

Contact

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

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.

Kim, Jin-Dong, Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1). pp. i180-i182, Oxford University Press, 2003.

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.