UIMA/U-Compare OpenNLP Sentence Detector

1. BASIC INFORMATION

Tool name

U-Compare Sentence Detector

Overview and purpose of the tool

This is a UIMA1 (Ferrucci et al., 2006) wrapper for the OpenNLP Sentence Detector tool. It splits English text into individual sentences.

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 tools3 are trained using machine-learning methods. The tool provided uses the pre-trained tagging 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 ULoader.class should be downloaded from http://u-compare.org/downloads/ULoader.class

Execution instructions

U-Compare is started by running ULoader.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 ULoader

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 purpose of the tool is to detect sentences in the text. An annotation is thus added to the CAS corresponding to each sentence in a document. 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 part of the output of the tool that is produced in in the U-Compare workbench. Each sentence detected 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)
Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 0.7 seconds.

4. LICENCES

The UIMA wrapper code, the underlying OpenNLP Sentence Detector 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

