U-Compare Cafetiere English Sentence Detector

1. BASIC INFORMATION

Tool name

U-Compare Cafetiere English sentence detector

Overview and purpose of the tool

The purpose of the tool is to detect sentence boundaries in English text. The tool is provided as a UIMA[1] (Ferrucci et al., 2006) component, specifically as Java archive (jar) file, which can be incorporated within any UIMA workflow. However, it is particularly designed use in the U-Compare text mining plaform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record), since the types of annotations it produces are compliant with the U-Compare.

A short description of the algorithm

The sentence detector uses a set of rules to break texts into sentences.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool can most easily be run using the U-Compare platform (see separate META-SHARE record). The only requirement to run U-Compare is for Java 6 to be installed.

For use in UIMA workflows outside of U-Compare, UIMA will need to be installed; see: http://uima.apache.org/

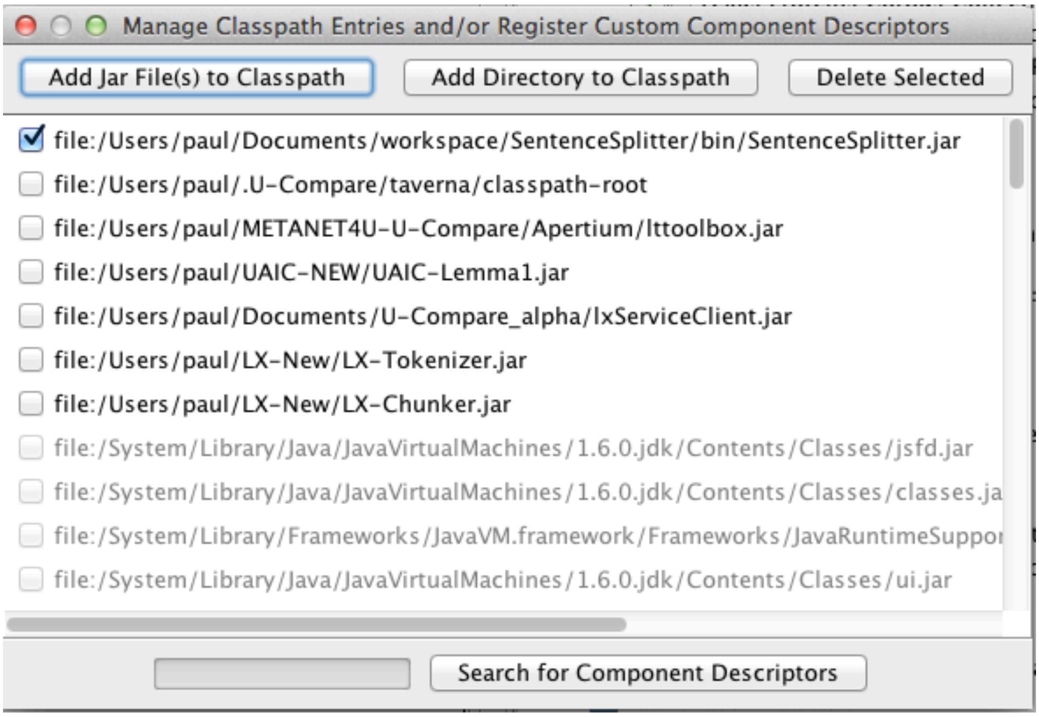
Installation

In order to run the sentence splitter in U-Compare, it must be imported into the system. Information about downloading and running U-Compare can be found in the documentation associated with the META-SHARE record for the U-Compare Workbench or the U-Compare website: http://nactem.ac.uk/ucompare/.

Importing the UIMA component (provided as the file CafetiereEngilshSentenceSplitter.jar) is carried out in U-Compare as follows:

1) From the “Library” menu in the U-Compare Workbench, choose the item “Register External Components (Edit Classpath)”. This will cause a window to appear that allows external components to be managed. It is shown in Figure 1.

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

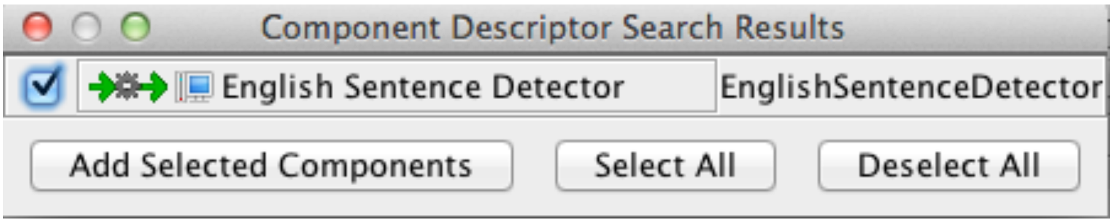

Manage Classpath Entries and/or Register Custom Component Descriptors
Add Jar File(s) to Classpath
Add Directory to Classpath
Delete Selected
Search for Component Descriptors

☑ file:/Users/paul/Documents/workspace/SentenceSplitter/bin/SentenceSplitter.jar
□ file:/Users/paul/.U-Compare/taverna/classpath-root
□ file:/Users/paul/METANET4U-U-Compare/Apertium/lttoolbox.jar
□ file:/Users/paul/UAIC-NEW/UAIC-Lemma1.jar
□ file:/Users/paul/Documents/U-Compare_alpha/lxServiceClient.jar
□ file:/Users/paul/LX-New/LX-Tokenizer.jar
□ file:/Users/paul/LX-New/LX-Chunker.jar
□ file:/System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Classes/jsfd.jar
□ file:/System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Classes/classes.ja
□ file:/System/Library/Frameworks/JavaVM.framework/Frameworks/JavaRuntimeSuppoı
□ file:/System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Classes/ui.jar

Figure 1: External component management window

2) Click on the button "Add Jar File(s) to Classpath", and browse to the location where the file "CafetiereEngilshSentenceSplitter.jar" has been saved, and click on "Open". The will cause the file to be displayed in the list of files in the external component manager window.

3) Ensure that box next to the file "CafetiereEngilshSentenceSplitter.jar" is checked in the external component manager window. Then, click the "Search for Component Descriptors" button. A "Component Descriptors Search Results" window will appear (Figure 2).



☑ English Sentence Detector EnglishSentenceDetector

Figure 2: Component Descriptor Search Results window

4) Check the box next to "English Sentence Detector", and click on the "Add Selected Components" button. The name of the component will then appear in the library on components on the right hand side of the main U-Compare workbench window, under "Custom components", and it can then be used in workflows.

Execution instructions

One imported into U-Compare, the sentence splitter can be used simply by dragging and dropping it into a workflow using the graphical user interface of the U-Compare workbench. Alternatively, it can be incorporated into other UIMA-based workflows, by following the documentation on the Apache UIMA site: http://uima.apache.org/

Input/Output data formats

Input data formats

The input is plain text document that has previously been read into the UIMA Common Analysis Stucture (CAS) via a UIMA collection reader component, i.e. this will normally be the first annotation tool that is run in a workflow/

Output data format

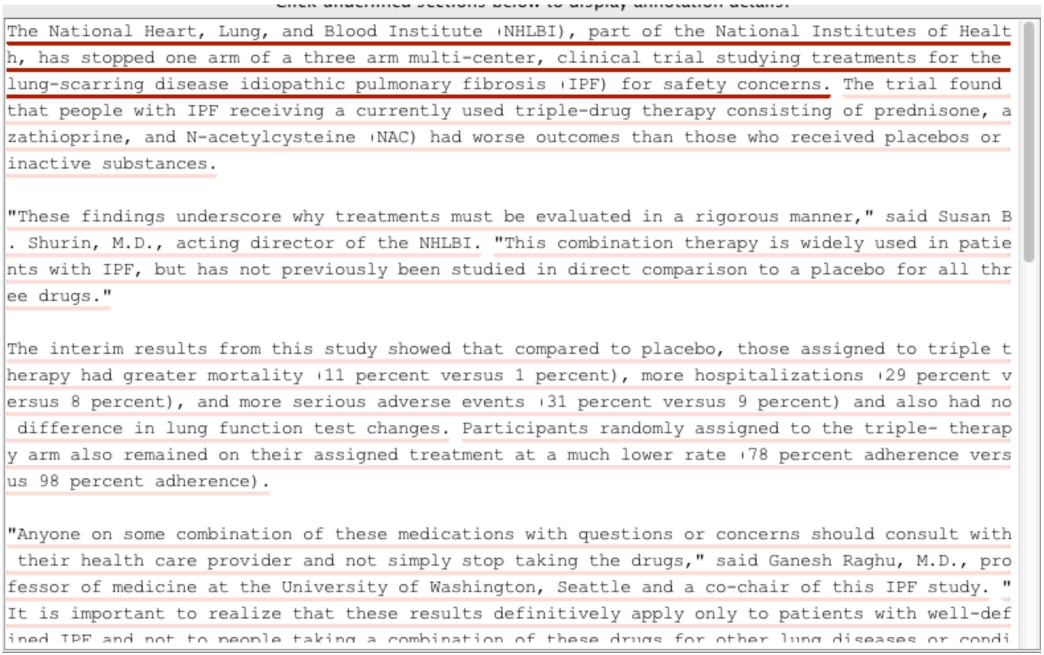
The tool detects the boundaries of sentences and adds an annotation of the type `org.u_compare.shared.syntactc.Sentence` (which is one of the types in the interoperable U-Compare type system) for each sentence in the document to the UIMA CAS.

Integration with external tools

As mentioned above, the tool can only be run within U-Compare framework, or more generally, within the UIMA framework.

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 Healt
h, 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, a
zathioprine, 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 patie
nts with IPF, but has not previously been studied in direct comparison to a placebo for all thr
ee drugs."

The interim results from this study showed that compared to placebo, those assigned to triple t
herapy had greater mortality (11 percent versus 1 percent), more hospitalizations (29 percent v
ersus 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- therap
y arm also remained on their assigned treatment at a much lower rate (78 percent adherence vers
us 98 percent adherence).

"Anyone on some combination of these medications with questions or concerns should consult with
 their health care provider and not simply stop taking the drugs," said Ganesh Raghu, M.D., pro
fessor of medicine at the University of Washington, Seattle and a co-chair of this IPF study. "
It is important to realize that these results definitively apply only to patients with well-def
ined IPF and not to people taking a combination of these drugs for other lung diseases or condi
```

Figure 1: Output of Cafetiere 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 30 milliseconds.

4. LICENCES

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

b) The underlying Cafetiere sentence splitter tool is licensed using the NaCTeM Software Licence Agreement (standard non-commercial use) – see “Cafetiere-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.

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.