UIMA/U-Compare GENIA Tokeniser (GENIA Tagger)

1. BASIC INFORMATION

Tool name

U-Compare GENIA Tokeniser (GENIA Tagger)

Overview and purpose of the tool

Tokenisation is one of the functionalities of the GENIA tagger, which additionally outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as MEDLINE abstracts.

The tool is 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

The tokenisation and POS tagging functionality is based on an algorithm described in Tsuruoka et al. (2005), which uses a cyclic dependency network (Toutanova et al, 2003) with maximum entropy modelling with inequality constraints. The tokenisation and POS tagging functionality was trained on a corpus containing newspaper articles (Wall Street Journal corpus), and the GENIA (Kim et al., 2003) and PennBioIE corpora (Kulick et al., 2003), both containing biomedical text.

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

[1] http://uima.apache.org/
[2] http://nactem.ac.uk/ucompare/

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
```

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 GENIA 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

One of the functionalities of the tool is to detect tokens in the text and assign parts-of-speech and base forms to them. An annotation is thus added to the CAS corresponding to each token in a document. Other annotations are also added by the GENIA tagger (e.g. named entity and chunk annotations), but we only focus on the token annotations here. 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)

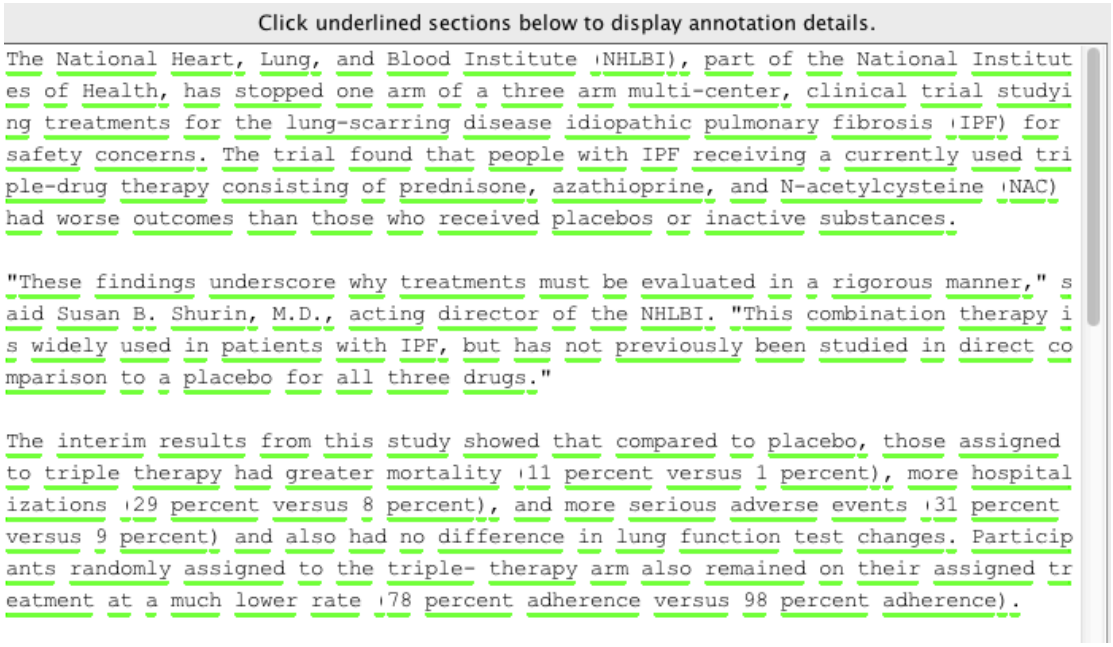

Figure 1: Output of the tokenisation functionality of the GENIA Sentence Detector in the U-Compare workbench

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 2.4 seconds.

4. LICENSES

a) The UIMA wrapper code is licensed using the NaCTeM Software Licence Agreement (standard non-commercial use) – see "GENIA-Tagger-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-Tagger-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.

Kim, J.-D, Ohta, T., Tateisi, Y. and Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1). pp. i180-i182, Oxford University Press, 2003.

Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A, Ungar L, Winters S, White P (2004). Integrated annotation for biomedical information extraction. Human Language Technology conference/North American Chapter of the Association for Computational Linguistics Annual Meeting, pp. 61-68.

Toutanova, K. and Klein, D. and Manning, C. D. and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of NAACL '03, pp 173- 180.

Tsuruoka, Y., Tateisi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S. and Tsujii, J.. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text. *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp 382--392, Springer-Verlag