UIMA/U-Compare OpenNLP POS Tagger

1. BASIC INFORMATION

Tool name

U-Compare OpenNLP POS Tagger

Overview and purpose of the tool

This is a UIMA[1] (Ferrucci et al., 2006) wrapper for the OpenNLP Tokenizer tool. It assigns part-of-speech tags to tokens in English text. The tagset used in from the Penn Treebank (Marcus et al., 1993).

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) 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 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 UCLoader.class should be downloaded from http://u-compare.org/downloads/UCLoader.class

Execution instructions

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

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 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 as input text that has been split into sentences and tokenised. Thus, the UIMA Common Analysis Structure (CAS) must contain both sentence and token annotations before this component is run. In a UIMA workflow, this could be achieved either by executing component(s) that perform sentence splitting and tokenisation prior to this component, or otherwise reading in a corpus of documents that already contains sentence and token annotations.

Output data format

The purpose of the tool is to detect tokens in the text. An annotation of type "POSToken" is thus added to the CAS corresponding to each token in a document. This type of annotation has a "posString" attribute to store the part-of-speech of the token. 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. The attributes of POSToken annotations are shown, consisting the of the beginning and end offsets of the token in the text, and the POS tag assigned to it. The sample text is taken the US National Library of Medicine website (http://www.nlm.nih.gov/databases/alerts/2011_nhlbi_ifp.html)

Covered Text	begin	end	posString
The	0	3	DT
National	4	12	NNP
Heart,	13	19	NNP
Lung,	20	25	NNP
and	26	29	CC
Blood	30	35	NNP
Institute	36	45	NNP
(NHLBI),	46	54	IN
part	55	59	NN
of	60	62	IN
the	63	66	DT
National	67	75	NNP
Institutes	76	86	NNPS
of	87	89	IN
Health,	90	97	NNP
has	98	101	VBZ
stopped	102	109	VBN
one	110	113	CD
arm	114	117	NN
of	118	120	IN
a	121	122	DT
three	123	128	CD
arm	129	132	NN
multi-center,	133	146	JJ
clinical	147	155	JJ
trial	156	161	NN
studying	162	170	VBG
treatments	171	181	NNS
for	182	185	IN
the	186	189	DT
lung-scarring	190	203	NN

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

Running the tool on the 4 KB text on a single core machine with 8 GB RAM takes around 3.4 seconds.

4. LICENCES

The UIMA wrapper code, the underlying OpenNLP POS Tagger 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

Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B. (1993), Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19(2), pp 313—330.

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.