UIMA/U-Compare STEPP Tagger

1. BASIC INFORMATION

Tool name

UIMA/U-Compare STEPP Tagger

Overview and purpose of the tool

Part-of-speech tagger tuned to biomedical text.

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) provides a graphical drag-and drop interface for the rapid creation of workflows.

A short description of the algorithm

The algorithm uses a combination of Conditional Random Fields (CRFs) (Lafferty et. al., 2001) and methods of maximum entropy (ME) tagging called two-phase ME tagging, which is based on an ME tagger introduced by Tsuruoka and Tsujii (2005). The combined model tagger achieves an accuracy of 97.20% when trained on the Penn TreeBank.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool is provided as a UIMA component wrapped around a web service. Thus, the tool must be run within the Apache UIMA framework. Alternatively, it can be run within the U-Compare framework. The component has been specifically designed to work in U-Compare workflows and is compliant with the U-Compare type system.

Installation

The tool is provided as an in-built component of the U-Compare workbench. However, it can also be used in other UIMA workflows. Since it is packaged as a UIMA component, no specific installation is required, following installation of the UIMA framework and/or U-Compare.

Execution instructions

The tool can be used within U-Compare simply be dragging and dropping it into a workflow using the graphical user interface of the U-Compare workbench.

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

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

Alternatively, it can be incorporated into other UIMA-based workflows, by following the documentation on the Apache UIMA site. Given that the UIMA component is implemented in Java, the tool is platform-independent.

Input/Output data formats

Input data formats

The input is plain text document that has previously been read into the UIMA Common Analysis Structure (CAS) via a UIMA collection reader component. There are two versions of the tool, which have different prerequisites:

1) STEPP Tagger - This version of the tool requires that both sentence and token annotations are present in the CAS prior to its execution. This can be achieved by including a sentence splitter tool and a tokeniser tool in the workflow, prior to the STEPP Tagger

1) STEPP Tagger with tokenization - This version of the tool only requires that sentence annotations are present in the CAS prior to its execution. This can be achieved by including a sentence splitter tool in the workflow, prior to the STEPP Tagger with tokenization

Output data format

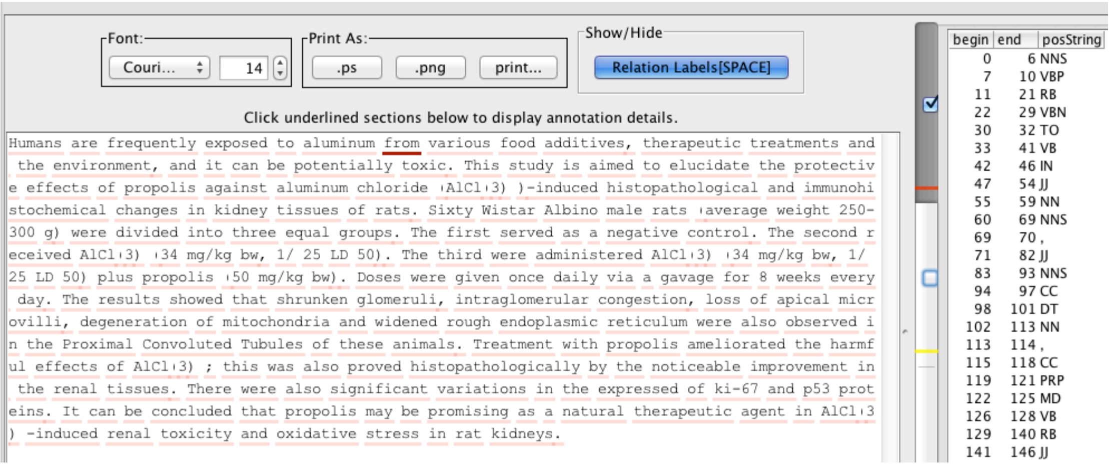
The tool creates a SteppToken annotation for each token in the text. This annotatins stores the part-of-speech assigned to the token. The "STEPP Tagger with tokenization" version will also split sentences into tokens.

Integration with external tools

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

3. CONTENT INFORMATION

Figure 1 shows the output of the tool in the U-Compare workbench. Each separate token is underlined in red. The POS tags assigned to each tag are shown in a tabular view. The sample text is taken from the the PubMed website (http://www.ncbi.nlm.nih.gov/pubmed/23172825)



Click underlined sections below to display annotation details.

```
Humans are frequently exposed to aluminum from various food additives, therapeutic treatments and
 the environment, and it can be potentially toxic. This study is aimed to elucidate the protectiv
e effects of propolis against aluminum chloride ıAlClı3) )-induced histopathological and immunohi
stochemical changes in kidney tissues of rats. Sixty Wistar Albino male rats ıaverage weight 250-
300 g) were divided into three equal groups. The first served as a negative control. The second r
eceived AlClı3) ı34 mg/kg bw, 1/ 25 LD 50). The third were administered AlClı3) ı34 mg/kg bw, 1/
25 LD 50) plus propolis ı50 mg/kg bw). Doses were given once daily via a gavage for 8 weeks every
 day. The results showed that shrunken glomeruli, intraglomerular congestion, loss of apical micr
ovilli, degeneration of mitochondria and widened rough endoplasmic reticulum were also observed i
n the Proximal Convoluted Tubules of these animals. Treatment with propolis ameliorated the harmf
ul effects of AlClı3) ; this was also proved histopathologically by the noticeable improvement in
 the renal tissues. There were also significant variations in the expressed of ki-67 and p53 prot
eins. It can be concluded that propolis may be promising as a natural therapeutic agent in AlClı3
) -induced renal toxicity and oxidative stress in rat kidneys.
```

begin	end	posString
0	6	NNS
7	10	VBP
11	21	RB
22	29	VBN
30	32	TO
33	41	VB
42	46	IN
47	54	JJ
55	59	NN
60	69	NNS
69	70	,
71	82	JJ
83	93	NNS
94	97	CC
98	101	DT
102	113	NN
113	114	,
115	118	CC
119	121	PRP
122	125	MD
126	128	VB
129	140	RB
141	146	JJ

Figure 1: Output of the U-Compare Stepp 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 870 milliseconds.

3. LICENCES

a) The UIMA wrapper code is licensed using the NaCTeM Software Licence Agreement (standard non-commercial use) – see "STEPP-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 Stepp Tagger web service is licensed using the NaCTeM Web Service Licence Agreement (standard non-commercial use)– see "STEPP-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.

4. ADMINISTRATIVE INFORMATION

Contact

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

5. REFERENCES

John Lafferty, AndrewMcCallum, and Fernando Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001, pages 282–289.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In Proceedings of HLT/EMNLP 2005. pages 467–474.