UIMA/U-Compare GENIA Tagger

1. BASIC INFORMATION

Tool name

UIMA/U-Compare Genia Tagger

Overview and purpose of the tool

The GENIA tagger analyzes English sentences and 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 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 used for the part-of-speech tagging is described in Tsuruoka et al. (2005). It is based on a bidirectional dependency network (Toutanova et al, 2003), but incorporates a simple, easiest-first strategy that is significantly more efficient than full bidirectional decoding. The easiest-first strategy performs comparably to full bidierectional inference, and experiments show that it consistently outperforms unidirectional inference methods. The part-of-speech part of the GENIA tagger is trained on the GENIA POS corpus (Tateisi & Tsujii, 2004). The GENIA tagger is trained not only on the Wall Street Journal corpus but also on the GENIA corpus and the PennBioIE corpus (Kulick et al, 2004).

The same algorithm is used to train the POS tagger was used to train the chunker. The Named Entity recogniser uses a sliding window with a maximum entropy classifier.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

In order to run U-Compare, Java 6 must be installed.

The component calls a web service. Therefore, internet access is required.

---

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

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

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

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 Tagger 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 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) GENIA 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 GENIA Tagger

1) GENIA 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 GENIA Tagger with Tokenization.

Output data format

The tool creates various types of annotations, and adds them to the CAS:

1) GeniaToken – a GeniaToken annotation is created for each token in the text, storing the POS tag and base form of the token

2) Chunk – corresponding to the syntactic chunks. Stores the chunk type.

3) GeniaRNA, GeniaDNA, GeniaProtein, GeniaCellType, GeniaCellLine- Named entity annotations corresponding to different types of biomedical entitites.

The "GENIA Tagger with tokenization" will also add basic Token annotations to the CAS.

Integration with external tools

As mentioned above, the tool can only be run as part of a workflow within the UIMA or U-Compare frameworks.

3. CONTENT INFORMATION

Figure 1 shows the output of the tool in the U-Compare workbench, with the various types of annotations underlined in different colours. The sample out is taken from the PubMed website (http://www.ncbi.nlm.nih.gov/pubmed/23150759)

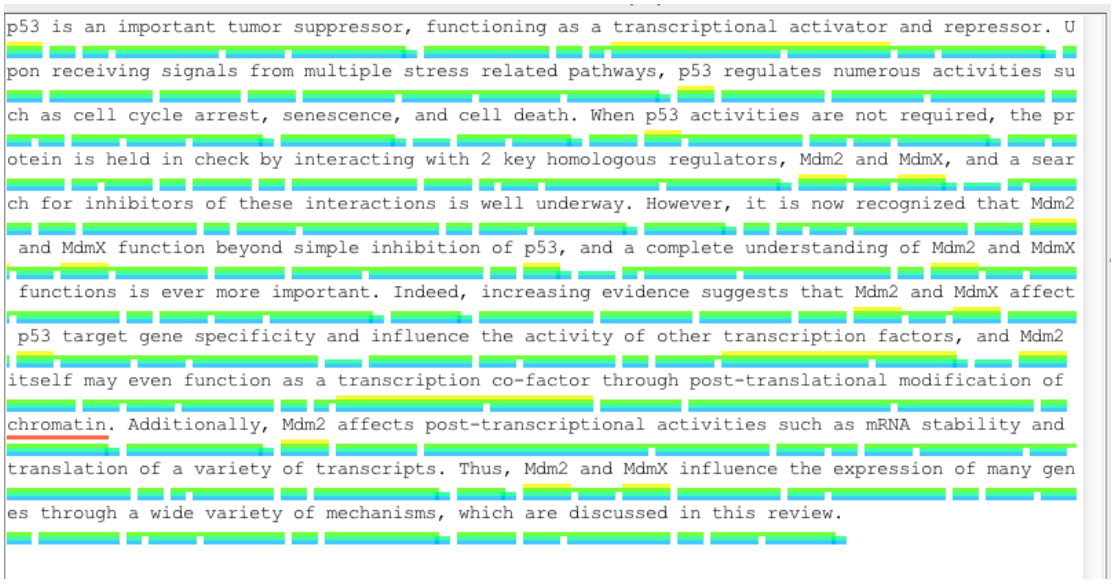

Figure 1: Output of the GENIA Tagger in the U-Compare workbench

3. LICENCES

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 GENIA Tagger web service that is called by the UIMA wrapper is licensed using the NaCTeM Web Service Licence Agreement (standard non-commercial use)– see "GENIA-Tagger-web-service-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 licences directory.

4. ADMINISTRATIVE INFORMATION

Contact

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

5. REFERENCES

S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein and L. Ungar (2004). Integrated Annotation for Biomedical Information Extraction,. In *Proceedings of the HLT/NAACL 2004 Workshop: Biolink 2004*, pp. 61-68.

Tateisi, Yuka and Jun'ichi Tsujii (2004). Part-of-Speech Annotation of Biology Research Abstracts. In *Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004)*. pp. 1267-1270

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