UIMA/U-Compare NEMine

1. BASIC INFORMATION

Tool name

UIMA/U-Compare NEMine

Overview and purpose of the tool

The purpose of the tool is to identify gene and protein names in 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 is described in Sasaki et al. (2008). The algorithm takes a two-step approach where sentences are first tokenised and tagged based on a biomedical dictionary that consists of general English words and about 1.3 million protein names. Then, a statistical sequence labelling step predicts protein names that are not listed in the dictionary and, at the same time, reduces false negatives in the POS/PROTEIN tagging results. A major benefit of the approach taken is that a user, rather than a system developer, can easily enhance the performance by augmenting the dictionary. The tool achieved a F-Score of 73.78 on the standard JNLPBA-2004 data set of MEDLINE abstracts with named entity annotations (Kim et al., 2004)

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

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.

---

[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 NEMine 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 a plain text document that has previously been read into the UIMA Common Analysis Structure (CAS) via a UIMA collection reader component. A prerequisite of this tool is that sentence annotations should already be present in the CAS. This can be achieved by including a sentence splitter tool in the workflow, which should be executed prior to NEMine being run.

Output data format

The tool finds gene and protein names in the text and adds corresponding annotations to the CAS.

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, with recognised protein names underlined in red. The sample out is taken from the PubMed website (http://www.ncbi.nlm.nih.gov/pubmed/23150759)

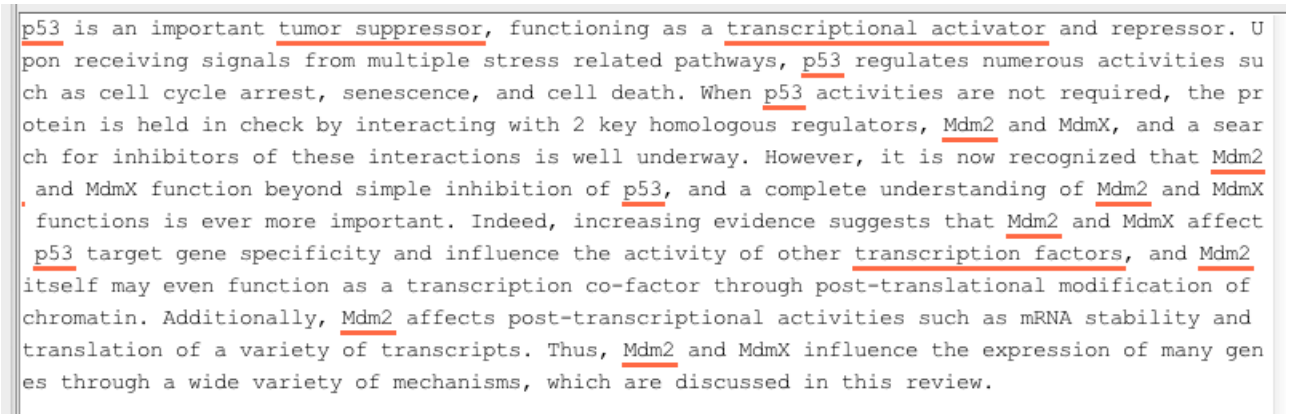

Figure 1: Output of NEMine in the U-Compare workbench

Running the tool on the 1 KB abstract on a single core machine with 8 GB RAM takes around 200 milliseconds.

4. LICENCES

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

c) The UIMA framework is licensed using the Apache licence. Please see "Apache-licence.txt" in the "licences" directory.

5. ADMINISTRATIVE INFORMATION

Contact

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

6. REFERENCES

Kim J-D, Ohta T, Tsuruoka Y, Tateisi Y (2004). Introduction to the Bio-Entity Recognition Task at JNLPBA. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pp. 70-75.

Sasaki, Y., Tsuruoka, Y., McNaught, J. and Ananiadou, S. (2008). How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics* 9(Suppl 11):S5.