GENIA Tagger

1. BASIC INFORMATION

Tool name

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.

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

The GENIA tagger can be run in the Unix operating system. Please note that a UIMA/U-Compare version of the GENIA tagger is also available, which allows the tagger to be run in text mining workflows, regardless of the operating system.

Installation

1) Download the tagger (source package for UNIX) from :

http://www.nactem.ac.uk/tsujii/GENIA/tagger/geniatagger-3.0.1.tar.gz

2) Expand the archive

```
> tar xvzf geniatagger.tar.gz
```

3) Make

```
> cd geniatagger/
> make
```

Also see the GENIA tagger page:

http://www.nactem.ac.uk/GENIA/tagger/

Execution instructions

Prepare a text file, containing one sentence per line, and then run the tagger as follows:

```
./geniatagger < RAWTEXT > TAGGEDTEXT
```

The tagger outputs the base forms, part-of-speech (POS) tags, chunk tags, and named entity (NE) tags in the following tab-separated format.

```
word1   base1   POStag1 chunktag1 NEtag1
word2   base2   POStag2 chunktag2 NEtag2
  :       :       :       :        :
```

Chunks and named entities are represented in the IOB2 format (B for BEGIN, I for INSIDE, and O for OUTSIDE).

Input/Output data formats

Input data formats

The input is plain text, with one sentence per line.

Output data format

The output is displayed with one token per line, with information output by the tagger separated by tabs. The information provided is as follows: surface form, base form, part-of-speech tag, chunk information and named entity information. See section 3 for an example.

Integration with external tools

N/A

3. CONTENT INFORMATION

An example of the output of the GENIA tagger is shown below.

```
Inhibition      Inhibition      NN      B-NP     O
of              of              IN      B-PP     O
NF-kappaB       NF-kappaB       NN      B-NP     B-protein
activation      activation      NN      I-NP     O
reversed        reverse         VBD     B-VP     O
the             the             DT      B-NP     O
anti-apoptotic  anti-apoptotic  JJ      I-NP     O
effect          effect          NN      I-NP     O
of              of              IN      B-PP     O
isochamaejasmin isochamaejasmin NN      B-NP     O
.               .               .       O        O
```

The chunk tags reveal that there are four noun phrases ("Inhibition", "NF-kappaB activation", "the anti-apoptotic effect", and "isochamaejasmin"). There is also one protein name (“NF-kappaB”).

3. LICENCES

The GENIA tagger is licensed using a proprietary non-commercial licence. Please see GENIA-LICENCE.txt file for details of the licence. For commercial use, please contact us using the details below.

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