STEPP Tagger

1. BASIC INFORMATION

Tool name

STEPP Tagger

Overview and purpose of the tool

Part-of-speech tagger tuned to biomedical text. Given plain text, sentences and tokens are identified, and tokens are assigned part-of-speech tags.

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).

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool is provided as a web service. Thus, the tool can be run from any computer connected to the internet.

Installation

No installation is required to run the web service. It is available at http://nactem001.mib.man.ac.uk:8080/UCompareDemo/SteppTaggerWithTokenisation

Execution instructions

The web service provides a demo form that allows the web service to be tested. An example of calling the web service from Java code is as follows:

```
String text = "Hello Mr. John Smith !";
String parameters = "text=" + URLEncoder.encode(text, "UTF-8") + "&mode=inline";
URL url = new URL ("http://nactem001.mib.man.ac.uk:8080/UCompareDemo/SteppTaggerWithTokenisation");
URLConnection connection = url.openConnection();
connection.setDoOutput(true);
OutputStreamWriter writer = new OutputStreamWriter(connection.getOutputStream());
writer.write(parameters);
```

```
writer.flush();
BufferedReader reader = new BufferedReader
(new InputStreamReader(connection.getInputStream(), "UTF-
8"));
String line;
while ((line = reader.readLine()) != null) {
    System.out.println(line);
}
```

Input/Output data formats

Input data formats

The input to the web service is plain text.

Output data format

The web service outputs XML. There is a `Sentence` element for each sentence identified in the input text, and a `SteppToken` element for each token identified. The `SteppToken` elements have 3 attributes, `begin`, `end` and `posString`, which store the beginning and end offsets of the token, and the part-of-speech tag, respectively.

Integration with external tools

As mentioned above, the web service can be called from program code, and so can be incorporated into applications straightforwardly.

3. CONTENT INFORMATION

The demo interface for the web service is shown in Figure 1. The interface provides some sample texts. Using the radio buttons above the text area, The user can choose to view the XML output, or the more user-friendly “inline annotations”, which provide an HTML visualisation of the annotations produced by the web service. Part of this visualisation is shown in Figure 2. The user can choose which type of output annotations to view using the check boxes.

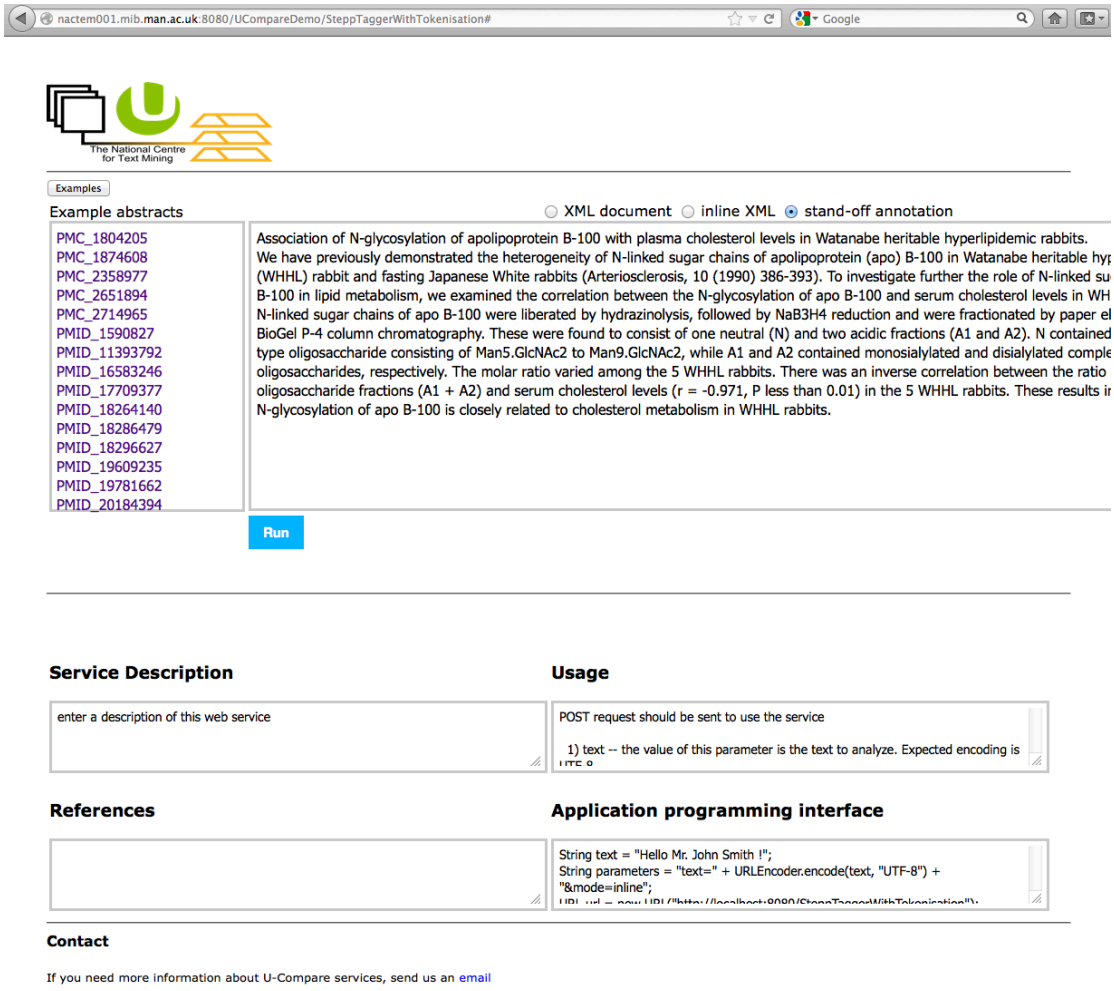
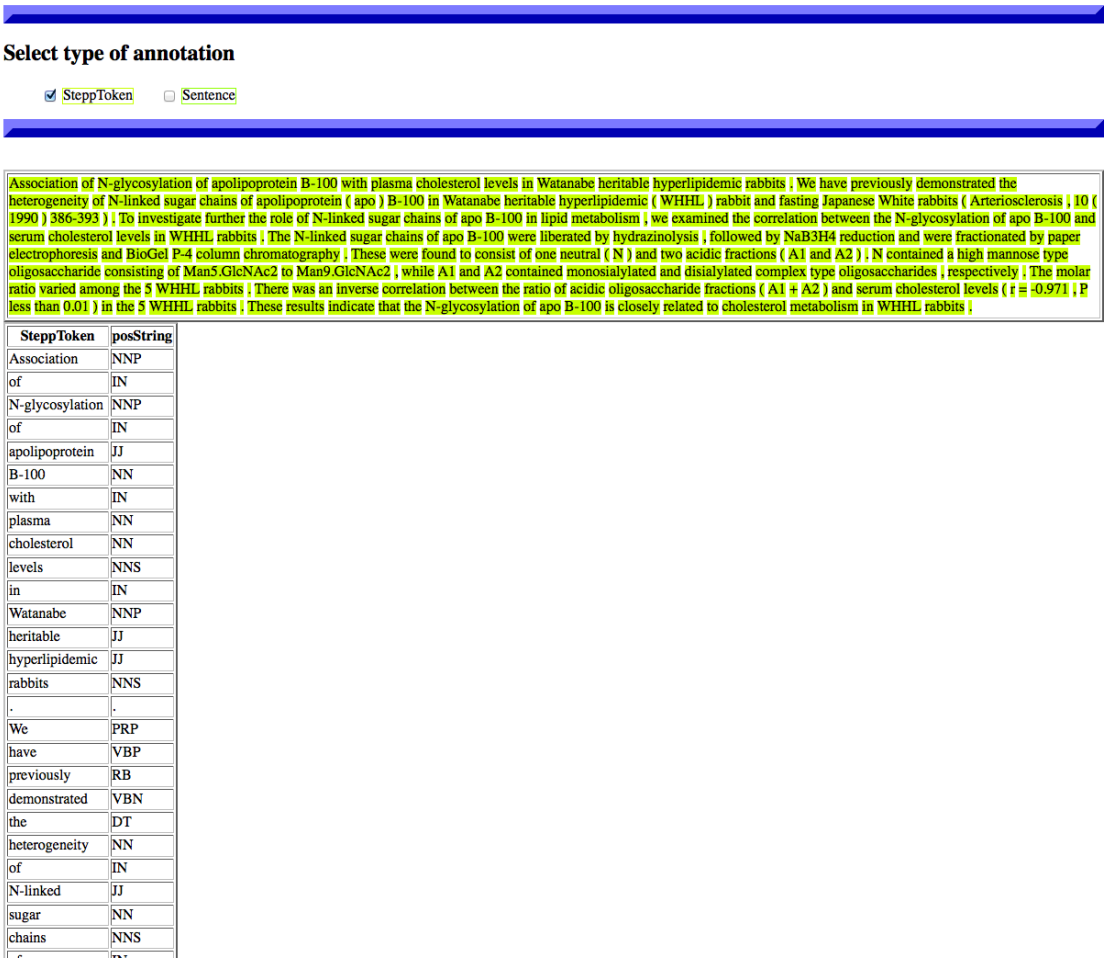


Figure 1: Demo interface for the STEPP tagger web service



Select type of annotation

☑ SteppToken ☐ Sentence

Association of N-glycosylation of apolipoprotein B-100 with plasma cholesterol levels in Watanabe heritable hyperlipidemic rabbits . We have previously demonstrated the heterogeneity of N-linked sugar chains of apolipoprotein (apo) B-100 in Watanabe heritable hyperlipidemic (WHHL) rabbit and fasting Japanese White rabbits (Arteriosclerosis , 10 (1990) 386-393) . To investigate further the role of N-linked sugar chains of apo B-100 in lipid metabolism , we examined the correlation between the N-glycosylation of apo B-100 and serum cholesterol levels in WHHL rabbits . The N-linked sugar chains of apo B-100 were liberated by hydrazinolysis , followed by NaB3H4 reduction and were fractionated by paper electrophoresis and BioGel P-4 column chromatography . These were found to consist of one neutral (N) and two acidic fractions (A1 and A2) . N contained a high mannose type oligosaccharide consisting of Man5.GlcNAc2 to Man9.GlcNAc2 , while A1 and A2 contained monosialylated and disialylated complex type oligosaccharides , respectively . The molar ratio varied among the 5 WHHL rabbits . There was an inverse correlation between the ratio of acidic oligosaccharide fractions (A1 + A2) and serum cholesterol levels (r = -0.971 , P less than 0.01) in the 5 WHHL rabbits . These results indicate that the N-glycosylation of apo B-100 is closely related to cholesterol metabolism in WHHL rabbits .

SteppToken	posString
Association	NNP
of	IN
N-glycosylation	NNP
of	IN
apolipoprotein	JJ
B-100	NN
with	IN
plasma	NN
cholesterol	NN
levels	NNS
in	IN
Watanabe	NNP
heritable	JJ
hyperlipidemic	JJ
rabbits	NNS
.	.
We	PRP
have	VBP
previously	RB
demonstrated	VBN
the	DT
heterogeneity	NN
of	IN
N-linked	JJ
sugar	NN
chains	NNS

Figure 1: Inline HTML visualisation of the output of the STEPP tagger, with Stepp Token annotations visualised. The table provides details of the annotations.

3. LICENCES

a) The 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.

b) The web service is dependent on the UIMA framework, which 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.