The CLAWS Web Tagger

Paul Rayson and Roger Garside
University of Lancaster

1 Introduction

Part-of-speech (POS) tagging, also called grammatical tagging, is the commonest form of corpus annotation, and was the first form of annotation to be developed by UCREL (University Centre for Computer Corpus Research on Language) at Lancaster. Our POS tagging software for English text, CLAWS (the Constituent Likelihood Automatic Word-tagging System), has been continuously developed since the early 1980s. The latest version of the tagger, CLAWS4, was used to POS tag 100 million words of the British National Corpus (BNC); see Garside (1996).

Several changes to the tagger were carried out during our work with the BNC. Tagset independence in the software was added as two tagsets were used in the BNC:

– a detailed tagset (C7) of 146 tags for a two million word sampler corpus, and
– a less refined tagset (C5) of 61 tags for the rest of the corpus.

We also moved to using SGML (Standard Generalised Mark-up Language) to represent features of a text; see Garside and Smith (1997: 109). Accented letters and other non-ASCII characters are now represented by SGML entities. For example, é for the lower case letter ‘e’ with an acute accent and £ for the pound sign.

We have now connected CLAWS to a web server to allow POS tagging over the internet.

2 Web Tagging

Readers wishing to access CLAWS on the web should point their internet browsers to:

http://www.comp.lancs.ac.uk/ucrel/claws/trial.html

This page contains a form with a text box into which up to three hundred words in normal orthography may be entered (or pasted). By clicking on the button within the form the text is submitted to CLAWS.

The text is passed via a CGI (common gateway interface) script written in Perl which runs on the webserver at Lancaster. This in turn saves the text into a file and passes control to a c-shell script which runs CLAWS over the file.

Within the webform, the user can select the tagset and output format for the text. If requested, the script converts the CLAWS vertical output (see Figure 1) into horizontal format (as in Figure 2), and maps the larger C7 tagset into C5. We now automatically process texts in the larger tagset and map to the smaller one as this results in more accurate output than processing in the smaller tagset; see Smith (1997: 141).

The CGI script then converts the text to a format suitable to pass back to the browser for the user to see.

0000003 010 The	AT
0000003 020 quick	[JJ/99] RR@/1 NN1%/0
0000003 030 brown	[JJ/93] NN1@/7 VV0%/0
0000003 040 fox	[NN1/100] VV0@/0
0000003 050 jumps	[VVZ/97] NN2@/3
0000003 060 over	[II/59] RP/41 NN1%/0 JJ%/0
0000003 070 the	AT
0000003 080 lazy	JJ
0000003 090 dog	[NN1/100] VV0%/0
0000003 091 .	.

Figure 1: vertical output from CLAWS

The_AT quick_JJ brown_JJ fox_NN1 jumps_VVZ over_II the_AT lazy_JJ dog_NN1 ._.

Figure 2: horizontal output from CLAWS

3 Usage

Usage of the CLAWS web tagger relies only on access to the internet site at Lancaster and a standard web browser such as Netscape Navigator. This allows the system to be accessed from PCs, Macs and UNIX machines. It has been employed from a PC lab in the Linguistics Department at Lancaster as part of an undergraduate course introducing grammar.

We first made the system available on the web in November 1997, and it has been accessed over 500 times in the three months since then.

References

Garside, Roger. 1996. The robust tagging of unrestricted text: the BNC experience. In J. Thomas and M. Short (eds) *Using corpora for language research: Studies in the Honour of Geoffrey Leech*, 167–180. London: Longman.

Garside, Roger and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, and A. McEnery (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, 102–121. London: Longman.

Smith, Nicholas. 1997. Improving a tagger. In R. Garside, G. Leech, and A. McEnery (eds) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, 137–150. London: Longman.