Abs-204

Sylwia Twardo (University of Warsaw)

Should we POS tag learner corpora?

Tagging errors in learner corpora is time-consuming and expensive. In order to make the task easier it would be useful to be able first to tag the parts of speech automatically. However, this may be feasible for learner corpora written by students who make few spelling mistakes. Removing texts with spelling mistakes from the analysis would blur the results. Hence it is worth-while to find a way of POS tagging corpora with spelling errors.

The present author POS tagged a learner corpus consisting of texts written by students at B1, B2 and C1 using CLAWS and analysed the errors in POS tagging. It was found that there were two kinds of them: those caused by the spelling mistakes in the texts and the errors made by the tagger and that both kinds of errors in tagging were systematic. The erroneous POS tags were tagged with the help of Spejd, a tool for rule based disambiguation.

References:

Buczyński A., Wawer A., (2008). Shallow parsing in sentiment analysis of product reviews. In: Proceedings of the Partial Parsing workshop at LREC 2008, pp. 14-18.

Buczyński A., Przepiórkowski A., (2008). Demo: An Open Source Tool for Partial Parsing and Morphosyntactic Disambiguation. In: Proceedings of LREC 2008.

Rayson, P. (2009). Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/