POS Tagging Errors in Learner Corpora

Sylwia TWARDO
Centre for Foreign Language Teaching
University of Warsaw
Warsaw, Poland
smtwardo@uw.edu.pl

This paper presents a tentative analysis of errors made by POS taggers (CLAWS and NLTK tagger) in a raw learner corpus and compares them to the results of POS tagging a learner corpus where words with the so-called non-word errors (Lee, 2009:12; Hovermale & Martin, 2008:3) were replaced with their corrected forms and manually POS tagged. The non-word errors have been classified into spelling errors (words where letters are switched, missing or added) and morphological errors (words which are composed of two correctly spelled parts, but the parts themselves are not correct, e.g. *tooked*). The aim of the analysis is to establish the degree to which correction of non-word errors reduces errors in POS tagging made by two different taggers. The results may be useful for creating a tool for semi-automatic correction of student errors.

The analysis was conducted with the use of the Learner Corpus of Student Examination work compiled at the Centre for Foreign Language Teaching at the University of Warsaw, which consists of 300,000 words. It comprises texts written by students at various University departments (except for the English Language Institute) for their English exams in 2003-2007 at levels B1, B2 and C1. Depending on the level, the texts are informal letters (B1) or opinion essays (B2 and C1), and vary in length from about 120-150 words (B1), 180-220 (B2), or 200-250 (C1). The students had to write these texts as part of a two-hour exam, which also included other tasks.

The corpus was first analysed as to the proportions and types of non-word errors and the length of the texts, and four clusters were selected with the use of simple K-means cluster analysis. They will be described in detail elsewhere, but, generally, Cluster 1 consists of relatively short texts with relatively few spelling and morphological errors per text, texts in Cluster 2 are of medium length on average and have a relatively high proportion of morphological errors, Cluster 3 has the longest texts and relatively few spelling and morphological errors, and Cluster 4 has texts of medium length and a relatively high proportion of spelling errors.

In order to make the current analysis more feasible, a sample of a learner corpus of student work was selected. Out of these 4 clusters two proportional representative samples of c.a. 10,000 words were randomly selected (the proportions were calculated as to the number of words, not of texts, hence there are more texts from Cluster 1 and fewer from Cluster 3). One sample came from Clusters 2 and 4 as they contained the greatest number of errors. The other sample was selected from Clusters 1 and 3.

The samples were POS tagged with the use of CLAWS (Rayson, 2009; UCREL CLAWS 7 Tagset) and the NLTK tagger (Bird, Klein, & Loper 2012) and the errors in POS tagging compared. Then the non-word errors were manually POS tagged (correctly) and correct equivalents of non-word errors were added. The version with corrected non-word errors was then POS tagged with the use of CLAWS and NLTK and the results were compared with the POS tags from the version with non-word errors.

References

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*. Available from: http://www.nltk.org/book. [1/05/2012].

Hovermale, D.J. & Martin S. (2008). 'Developing an Annotation Scheme for ELL Spelling Errors.' Department of Linguistics, The Ohio State University, Columbus, OH. Available from: http://www.ling.ohio-state.edu/~scott/publications/Hovermale-Martin-MCLC05-2008.pdf. [10/12/2010].

Lee, J.S.Y. (2009). *Automatic Correction of Grammatical Errors in Non-native English Text.* Unpublished PhD thesis. Cambridge, MA.: MIT. Available from: http://groups.csail.mit.edu/sls/publications/2009/Thesis_Lee.pdf. [10/12/2010].

Rayson, P. (2009). 'Wmatrix: a web-based corpus processing environment.' *Computing UCREL CLAWS7 Tagset*. Available from: http://ucrel.lancs.ac.uk/claws7tags.html. [10/12/2011].

UCREL CLAWS7 Tagset. Available from: http://ucrel.lancs.ac.uk/claws7tags.html. [10/12/2011].