NLP-based measurement of text quality for learner writing: Relationship between text similarity and proficiency

Boo Kyung Jung (University of Pittsburgh), Gyu-Ho Shin (Palacky University Olomouc), & Seongmin Mun (Université Paris Nanterre)

The present study investigates the degree to which Natural Language Processing (NLP) techniques for text quality measurement address proficiency of L2-Korean learners. With the recent development of NLP techniques, a number of L2 studies utilize them to analyze learner corpus automatically (Meurers, 2015). One area in learner corpus research is text quality, which concerns semantic-pragmatic aspects of language use to affect overall text quality. Despite increasing interests in employing NLP techniques, little attention has been paid to how similarly/differently each technique reveals constructs of L2 learners (e.g., proficiency). In addition, NLP-based L2 research is heavily biased toward L2 English. Against this background, we investigate how text similarity scores (as a proxy for text quality) calculated by NLP techniques representative of topic modelling (LSA, LDA) and word embedding (Word2Vec, BERT) explain proficiency.

We recruited 36 Mandarin-speaking learners of Korean and 10 native speakers of Korean. Each participant was asked to write argumentative essays on two topics. Only content words in each essay, without correction of typos and errors, were considered in the actual analysis. Learner essays (individually) and native speakers' essays (as a whole) were converted to vectors by using each technique; similarity scores of learner essays (relative to the native speaker writing) were calculated through cosine similarity. We then conducted linear regressions, with similarity scores of each technique as an independent variable and with proficiency scores (measured separately through the Korean C-test; Lee-Ellis, 2009) as a dependent variable. We also grouped the learner essays (by topic) into two groups—highest and lowest, each of which consisted of seven essays classified by each technique—to see if there are any group differences in the performance of each technique.

Results from the linear regression analysis revealed asymmetric degrees of model performance, which was also sensitive to the topics, in explaining proficiency scores: Word2Vec yielded significance in both topics; LSA and BERT showed significance only in Topic 2; LDA showed no significance in both topics, producing extreme similarity values (either 0 or 1). While the three techniques (LSA, Word2Vec, BERT) showed significance, all of them fell short of capturing individual variations involving learner writing, by classifying some participants into the two learner groups at the same time depending on essay topics. It was also found that LSA and Word2Vec performed better than LDA and BERT in general for the task of categorizing the essays from the same participants into the same learner groups, either highest or lowest, across the essay topics.

In sum, these results suggest that the degree that each NLP technique explains learner constructs (proficiency in this study) was asymmetrical and sensitive to essay topics (and particularly to word use such as repetitions of keywords). Given the recent trend that NLP techniques are widely used in learner corpus research, this study's findings call for a need for researchers to pay close attention to the specifics of NLP techniques for measuring text quality, in consideration of learner language characteristics (e.g., simple/short sentences, errors) and essay topics.

References

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing, 26*(2), 245–274.

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Eds.). *The Cambridge handbook of learner corpus research* (pp. 537–566). Cambridge: Cambridge University Press.