Multidimensional Vector Distance as Measure of Text Level Differences

Xiaobin Chen Detmar Meurers

Dec. 2015

1 Abstract [1]

Reading texts of appropriate levels enables language learners to practice being competent readers and motivates them to read more (Milone & Biemiller, 2014). A major concern for language teachers is thus on how to match students' reading abilities with the texts' complexity levels. Readability assessment provides a way to measure text complexity and assign text levels. The process has been automatized with natural language processing and machine learning tools, the outcome of which are usually grade/age level estimates of the reading texts, because the tools are built on classification models trained with features extracted from the text and their gold-standard levels provided by expert teachers or publishers of the text.

Despite the tools' highly accurate estimates, their application in actual reading system design is limited by the lack of a method to link the learner's proficiency to the text's readability level. The present study proposed using multidimensional vector distance as a measure of text level differences and using it to link learner proficiency and text complexity levels. Text features used in (Vajjala & Meurers, 2012) were extracted from both learner texts and reading texts to form vectors representing the students' proficiency and the reading text's complexity level. Our hypothesis was that vector distances were correlated with level differences, i.e., greater level differences would result in greater vector distance.

The hypothesis was tested with two data sets, one from Wang and Wang's (2015) continuation writing experiments and the other involved 30 randomly selected articles from the educational website Newsela (`www.newsela.com`). In Wang and Wang's (2015) study, 48 EFL students were required to continue writing stories in English after reading the beginning stories whose endings were ommited. It was found that after reading an English version of a story, students made less mistakes in their writings than after reading a Chinese version of it. The authors attributed the results to the alignment effect of language learning,

[1] This abstract was presented at the workshop *Studying Language Learning: From the Laboratory to the Classroom* held on Dec. 10–11th, 2015 in Tübingen University.

stating that the writings after reading the English story input had an improved quality and is closer to the level of the input texts. Using this data set, we tested the hypothesis that text vector distances between writings after reading Chinese input and the input text are greater than those between the English input and the writen output. The results with paired sample t-test confirmed this hypothesis ($T(47) = 3.35,\ p \leq .001$).

The 30 articles, each in 5 different reading levels, from Newsela were used to test the hypothesis that greater level gaps (e.g., the distance between levels 1 and 5 in comparison to that between levels 1 and 2) are associated with greater vector distances. One-way ANOVA and post-hoc Tukey test results confirmed the hypothesis ($F(3, 296) = 403.1,\ p < .001$). With the increase of level differences, larger vector distances were observed.

This study confirmed that it is valid to use distances between text feature vectors to measure the texts' reading level differences. It is also a valid method to link reading proficiency with text readability, hence a potentially viable method for automatic individualized reading text assignment for language development.

References

Milone, M., & Biemiller, A. (2014). *The development of ATOS: The Renaissance readability formula* (Tech. Rep.). Wisconsin Rapids: Renaissance Learning.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using nlp.*

Wang, C., & Wang, M. (2015). Effect of alignment on l2 written production. *Applied Linguistics*, *36*(5).