Reading Level Assessment for Literary and Expository Texts

Kathleen M. Sheehan (ksheehan@ets.org)
Irene Kostin (ikostin@ets.org)
Yoko Futagi (yfutagi@ets.org)
Educational Testing Service, MS 12-R, Rosedale Road
Princeton, NJ 08541 USA

Keywords: reading level assessment, readability, genre.

Introduction

Teachers and test developers routinely use automated text analysis systems to select reading materials that are challenging yet not overly difficult for the targeted population of readers. While such systems have traditionally relied on shallow linguistic features such as average sentence length and average word length, researchers have recently begun to explore more theoretically meaningful aspects of text variation. For example, Petersen and Ostendorf (2006) evaluated construct-relevant grammatical features such as the average parse tree height; McNamara, Ozuru, Graesser & Louwerse (2006) evaluated automated coreference indices such as the degree of sentence-to-sentence overlap; and Heilman, Collins-Thompson, Callan & Eskenazi (2007) evaluated additional construct-relevant grammatical features such as passive voice, relative clauses and modals.

Genre-related effects have also been evaluated. For example, researchers at the School Renaissance Institute (SRI, 2000) noted that the widely-used Flesch-Kincaid Grade Level score (F-K) tends to overpredict the difficulty of literary texts while simultaneously underpredicting the difficulty of expository texts.

The D-Tree System

This paper describes an automated reading level assessment system targeted at readers in grades 3 through 12. This new system, called D-Tree, builds on the research summarized above by (a) providing distinct models for literary and expository texts; and (b) using a Factor Analysis (FA) modeled after that reported in Biber (1988) to incorporate a large number of construct-relevant text features. The FA was implemented on a 13 million word corpus constructed from 12,854 texts targeted at students in grades 3 through 12. (Sheehan, Kostin, Futagi & Sabatini, 2006). It yielded nine dimension scores defined in terms of 50 distinct features. A tree-based regression analysis, implemented on a second corpus of 187 literary, and 156 expository passages, was then used to develop two prediction models: one for literary texts and one for expository texts. The nine dimension scores defined via the FA served as the candidate predictors considered in each model. The resulting system is evaluated in Table 1. Two types of evaluations are presented: the correlation between D-Tree's predictions of text grade level and corresponding human ratings, and a measure of the D-Tree/human bias estimated via the mean model residual. The results obtained when these same measures are calculated for the F-K score are also shown. The results confirm that D-Tree's predictions are both unbiased and highly correlated with human ratings. The F-K score, by contrast, yielded somewhat lower correlations while generally replicating SRI's results, i.e., the difficulty of the literary texts was overpredicted (by an average of 0.66 grade levels) and the difficulty of the expository texts was underpredicted (by an average of 1.35 grade levels.)

Table 1: Correlation Between Human and. Automated Assessments of Text Grade Level

Model	Genre	Correlation	Bias
F-K	Literary	0.69	+ 0.66
F-K	Expository	0.77	- 1.35
D-Tree	Literary	0.79[a]	0.00[a]
D-Tree	Expository	0.83[a]	0.00[a]

a. Estimated via 10-fold cross validation

Acknowledgments

This research was jointly funded by ETS and by a grant from the Institute of Education Sciences (R305G040065).

References

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Heilman, M.J., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2007). *Combining lexical and grammatical features to improve readability measures for first and second language texts.* Proceedings of the HLT Conference, Rochester, NY.

McNamara, D., Ozuru, Y., Greasser, A., & Louwerse, M. (2006). Validating coh-Metrix. In R. Sun & N. Miyake (Eds.), Proceedings of the 28th Annual Conference of the cognitive Science Society, Mahwah, NJ:Erlbaum.

School Renaissance Institute (2000). The ATOS readability formula for books and how it compares to other formulas. (Technical Report). Madison, WI: SRI.

Petersen, S.E. & Ostendorf, M. (2006). *A machine learning approach to reading level assessment.* University of Washington CSE Technical Report.

Sheehan, K.M., Kostin, I., Futagi, Y. & Sabatini, J. (2006). *Measuring the prevalence of spoken language structures in printed text: An approach for improving automated predictions of text difficulty.* Presented at the Society for Text and Discourse Conference, Minneapolis, MN.