Assessing Text Difficulty Development in ELT Textbooks Series Using N-gram Language Models based on BNC

Alvin Cheng-Hsien Chen (National Taiwan Normal University, Taiwan)

This study aims to explore the possibilities of developing a corpus-based evaluation system of text difficulty development for English Language Teaching (ELT) materials by drawing insights from the N-gram-based language models used in natural language processing. ELT materials evaluation has paid little attention to the appropriateness of the text difficulty development in a textbook series (Ghorbani, 2011; Mukundan & Ahour, 2010; Tomlinson, 2012; Tsagari & Sifakis, 2014). The growing availability of machine-readable large corpora has facilitated the computation of frequency ratings for words, thus offering objective statistics for lexical sophistication in a given text (Crossley, Cobb, & McNamara, 2013). A heuristic in using the frequency ratings of words from large corpora is: words of lower frequency tend to be more "sophisticated". Based on this principle, we aim to propose a new method for assessing the progression of lexical sophistication in ELT materials, or text difficulty in general, which will be argued to be a favorable measurement compared to the band-based frequency approach in the research paradigm of Lexical Frequency Profile (Laufer and Nation, 1995). Specifically, we addressed three research questions: (1) How can N-gram-based language models in NLP be applied to the field of computational text analysis and to the measurement of text difficulty? (2) Can an N-gram-based language model properly characterize the text difficulties development of an ELT textbook series and assess the developmental trends and transitional gaps of the text difficulty progression? (3) Does the N-gram-based method provide a better assessment of text difficulty development than the band-based frequency method in LFP?

For N-gram-based method, we built a trigram language model on British National Corpus (XML Edition), using self-developed scripts, by converting the raw frequencies of all the trigrams into probabilities via normalization. In case of unseen events, we adopted the state-of-art smoothing algorithm — Kneser-Ney Smoothing — to estimate the probabilities of the unseen (Jurafsky & Martin, 2008). Then the generated Trigram Language Model was used to compute the *log probability* of each sentence in a textbook volume, serving as a measure of degrees in lexical sophistication under the condition that the language model is representative enough. That is, each volume of the ELT textbook series was quantitatively measured by a series of log probabilities. As for the band-based LFP method, we used the BNC unlemmatized word lists provided by Adam Kilgariff as our base list and generated the top thirteen 1000-word lists according to the procedure described in Chen (2016). Following the LFP framework, the contents of each volume in the ELT textbook series was described by the coverage rates of each 1000-word list across thirteen frequency bands. Therefore, the text difficulty of a given text was represented by two different sets of metrics: Log Probabilities (N-gram method) and coverage rates (band-based method). Two experiments were conducted in this study.

Our first experiment adopted the bottom-up clustering-based method to identify the developmental trends and transitional gaps of text difficulty in a six-volume textbook series, using the algorithm of Variability Neighboring-joining Clustering, proposed by Gries and Hilpert (2008). We collected a corpus of senior high school textbooks (SHSTC), including three major officially-approved versions of textbook series used in the compulsory English curriculum in Taiwan. Each textbook series has six volumes targeted for use in six different semesters (in three years). The VNC proceeded as follows. The first step was to evaluate the distance between each pair of neighboring volumes (e.g. V1 vs. V2, V2 vs. V3, …, V5

vs. V6) and the second step was to construct a hierarchical tree by merging volumes of the minimal distance. In terms of the distance metric, we used the *p*-values from Kruskal Wallis Tests for the log probabilities from the N-gram method and Pearson's correlation coefficients for the coverage rates from the band-based method. In each step of merging, all the pairwise distance matrix was computed and the two neighboring volumes/super-clusters of the minimal distance were merged into one super-cluster. This merging proceeded iteratively until all the six volumes were merged into one node. After the dendrogram was created by the VNC, we created the plot of within-cluster variance (Kaufman & Rousseeuw, 2005) to determine the number of stages in the text difficulty development. The general principle is that the within-cluster variance decreases as the number of clusters increases. The local minimum (i.e. an elbow point) of this plot provides a possible optimal number of transitional gaps in our dataset.

In our second experiment, we evaluated the performance of the two metrics from N-gram and band-based methods, by applying them to analyze the text difficulty of four distinct types of ELT materials. In addition to the SHSTC, we further collected three types of ELT materials: (1) Junior High School Textbooks Corpus (JHSTC), (2) College Entrance Examination Corpus (CEEC), (3) Teacher Recruitment Examination Corpus (TREC). The JHSTC contained the official textbook materials used in the junior high school in Taiwan. The CEEC included all the entrance examination papers administered to the senior-high-school graduates as an official qualifying exam for their college/university admission. The TREC included examination papers administered as qualifying exams for prospective English teacher recruitment in the secondary education in Taiwan. We assume that the difficulty levels should be: JHSTC < SHSTC < CEEC < TREC, where the junior-high school textbooks (i.e. JHSTC) are expected to be the least sophisticated texts and the teacher recruitment examinations are expected to be the most challenging texts from an L2 learner's perspective. Under this assumption, we argue that a better metric of text difficulty should be the one that can maximize the difficulty levels among these four types of dataset. For each type of corpus, we computed their N-gram-based metric (i.e. log probabilities of sentences in each corpus) as well as band-based metric (i.e. LFP coverage rates in each corpus) as indices for text difficulty (or lexical sophistication). We then conducted a variance-based statistical test (*F*-test) to see which metrics would generate the maximal between-group variances. That is, the metric that can maximize the differences among these four corpora may better distinguish the variation of the lexical sophistication among these four text difficulty levels.

Our results show that the N-gram-based estimation of lexical sophistication yields comparable results to the prediction by the band-based method and the former generates a better metric that can maximize the variation of the text difficulties among the four types of ELT materials we collected. Furthermore, we argue that N-gram-based method is more favorable in several aspects. First, an N-gram language model does not have to make an arbitrary assumption with respect to the size of the frequency bands (e.g. 1000-word), thus running less risk of over-estimating the vocabulary level of the text. Second, the language model provides not only more fine-grained and precise estimates of probabilities for words but also probabilistic estimates for word combinations (e.g. multiword expressions), thus taking into account considerable degrees of contextual sophistication of collocation/colligation patterns. Third, the N-gram language model provides a more holistic measure for lexical sophistication by considering the low-frequency patterns in different frequency scales. Most importantly, these probabilities were based on a large-scale representative corpus, which provides a common ground for cross-cultural comparative evaluations of ELT materials, thus shedding light on several pedagogical implications for EFL/ESL learners and teachers as well as ELT textbook developers.

References

Chen, A. C.-H. (2016). A critical evaluation of text difficulty development in ELT textbook series: A corpus-based approach using variability neighbor clustering. *System, 58*, 64-81.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System, 41*, 965-981.

Ghorbani, M. R. (2011). Quantification and graphic representation of EFL textbook evaluation results. *Theory and Practice in Language Studies, 1*(5), 511-520.

Gries, S. T., & Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora, 3*(1), 59-81.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition* (2nd edn ed.). Upper Saddle River, NJ: Prentice Hall.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis* (2nd edn ed.). Hoboken, NJ: Wiley.

Mukundan, J., & Ahour, T. (2010). A review of textbook evaluation checklists across four decades (1970–2008). In B. Tomlinson & H. Masuhara (Eds.), *Research for materials development in language learning: Evidence for best practice* (pp. 336-352). London: Continuum.

Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching, 45*(02), 143-179.

Tsagari, D., & Sifakis, N. C. (2014). EFL course book evaluation in Greek primary schools: Views from teachers and authors. *System, 45*, 211-226.