Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Detecting Fine-Grained Syntactic Features for Predicting Japanese EFL Learners' Writing Proficiency

Takeshi KATO[a*], Yuichi ONO[b]
[a*]*Doctoral Program in Literature and Linguistics, University of Tsukuba, Japan*
[b]*Faculty of Humanities and Social Sciences, University of Tsukuba, Japan*
*takeshi.kato1994@gmail.com

Abstract: In the field of foreign language writing research, linguistic features (e.g., mean length of sentence, type-token ratio) that appear in learners' performance have been utilized to gauge learners' development. Among them, syntactic aspect of linguistic structure has prominently been investigated. While there are wide variety of features, however, it is not clear what kind of features should be implemented to improve the quality of writing assessment. This study aims at detecting fine-grained syntactic features for predicting Japanese English-as-a-foreign-language (EFL) learners' writing proficiency. In the analysis, we used 5,000 argumentative essays written by Japanese learners of English, which are assigned five proficiency levels. A total of 78 fine-grained syntactic features are computed from the essays with TAASSC (Kyle, 2016) and employed as predictors of proficiency levels in a random forest classifier. The results suggest that noun phrase elaboration, use of modal, use of passive voice, and verb based syntactic knowledge contribute to the prediction of the proficiency levels.

Keywords: automated essay scoring, syntactic feature, second language writing assessment, random forest

1. Introduction

Since the late 1990s, second language (L2) acquisition and education researchers have focused on the relationship between L2 learners' proficiency and textual features which appear in learners' written and spoken performance (e.g., Wolfe-Quintero, Inagaki, & Kim, 1998). Complexity, Accuracy, and Fluency (CAF) triad has been the most general framework for the purpose of analyzing L2 leaners' development and numerous linguistic features to gauge the three concepts have been proposed (see Housen, Kuiken, & Vedder, 2012). Among them, syntactic complexity, which is a sub-concept of Complexity, has prominently been employed in a great number of L2 writing and speaking studies.

Syntactic complexity, defined as the variety and degree of sophistication of the syntactic structures used, has been measured multidimensionally with variety of linguistic features (Norris & Ortega, 2009; Ortega 2003). Such measuring methods are supported by automated analyzers, which enable us to analyze texts by multiple syntactic levels: sentential, clausal, and phrasal levels (e.g., Biber, Gray, & Poonpon, 2011; Kyle, 2016; Lu, 2010; McNamara, Graesser, McCarthy, & Cai, 2014).

On one hand, thanks to the advance of natural language processing technique, detailed syntactic analyses with diverse linguistic features in automated essay scoring. On the other hand, it has become difficult to determine which factors among syntactic complexity contribute to the prediction of learners' proficiency, due to the proliferation of proposed diverse features as the research proceeds further. For instance, although recent studies have attempted to predict learners' proficiency or writing quality with those automated syntactic features using linear regression, many such features have been excluded from predictive models due to merely statistical problems (e.g., multiple collinearity), resulting in poor accuracy of the models (e.g., Kyle & Crossley, 2017, 2018). This treatment causes not only decrease in explanatory power of the predictive model but loss of representativeness of the construct (i.e., syntactic complexity). In order to construct a more accurate model, it is necessary to identify factors which contribute to predicting learners' proficiency level without a priori feature deletion. Working on this task can lead us to the construction of the writing support system for ESL learners.

Therefore, this study attempts to solve the above-mentioned issue. In so doing, we employ random forest algorithm (Breiman, 2001) as a prediction method and detect syntactic features that contribute to predicting learners' proficiency among all relevant features.

2. Methodology

2.1 Dataset

This study employs the essay data from a large scale standardized English proficiency test (EIKEN foundation of Japan) held in Japan. All the test takers were Japanese EFL learners and each essay was assigned one of five proficiency levels: from A1 to C1 level of CEFR. For the subsequent analysis, 1,000 essays were randomly selected from each proficiency level; the dataset contained 5,000 argumentative essays.

2.2 Automated Analyzer and Relevant Features

TAASSC (Kyle, 2016) was utilized for automated analysis of syntactic sophistication and complexity and relevant syntactic features were computed from the essays. This analyzer deals with three types of syntactic features related to syntactic complexity, which overwhelm the defects of conventional features: (i) fine-grained noun phrase complexity features, (ii) fine-grained clausal complexity features, and (iii) syntactic variation and sophistication features. The first type of features quantifies noun phrase elaboration (e.g., prepositional modification of noun phrases). The second type counts grammatical elements per clause (e.g., adverbial modifiers per clause). The third type operationalizes syntactic structural variety and sophistication utilizing Corpus of Contemporary American English (COCA) as a reference corpus. A total of 78 relevant syntactic features were selected from these types of features.

2.3 Implementation of Random Forest Algorithm

The analysis was conducted by using the statistical programming language R (R Core Team, 2019). The package randomForest (Liaw & Wiener, 2002) was used to implement the algorithm. We constructed a predictive model where 78 syntactic features as independent variables and the five proficiency levels as dependent variables. In learning, the number of independent variables subject to random sampling was set to the square root of the number of independent variables and the number of trees was set to 500.

3. Result and Discussion

As a result of cross-validation by OOB (out-of-bag), the prediction accuracy of the model was 50.34%, which largely exceeds that of previously proposed multiple linear regression models: 14.2% (Kyle & Crossley, 2017) and 20.3% (Kyle & Crossley, 2018).

Figure 1 shows the top 10 of the 78 independent variables in descending order of their contribution to classification, according to their mean decrease Gini coefficients. The first, second, third and seventh features are related to noun phrase elaboration, and in particular, noun phrase modification by prepositional phrases and adjectives contributed to the classification. The fourth through sixth features are the ones of clause level and related to the use of modal auxiliaries and passive voice. The features from the eighth to the tenth are features related to the association strength between a matrix verb and syntactic structure of a sentence, and are operationalization of knowledge related to the sentence structure based on a verb.

The high performances of the fine-grained noun phrase complexity features observed above support Biber et al.'s (2011) finding that noun phrase modification plays an important role in writing development. Additionally, clausal and syntactic structural features' contribution implies that the necessity of investigating grammatical variety and sophistication in assessment of syntactic complexity and L2 writing.

The result suggests that the proposed method is more promising than the conventional ones, but there is much room for improvement particularly in generalizing the current model. For future research, it is necessary to compare the result obtained in this study with that of other classification methods (e.g.,

boosting), and analyze different types of essays (e.g., narrative) and data by English learners whose native language is other than Japanese.

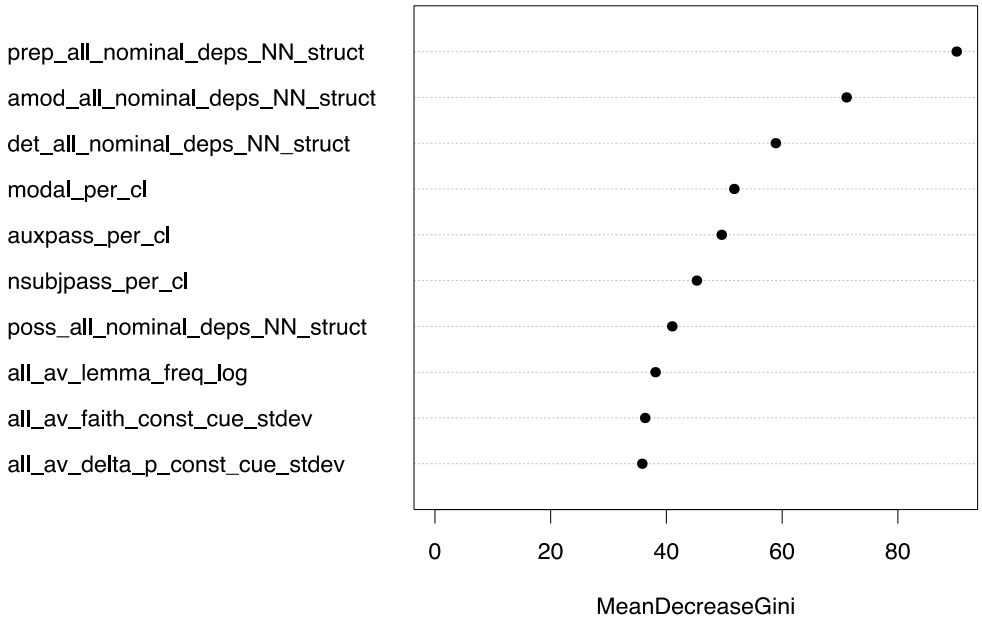

Figure 1. The top 10 contributing features.

Acknowledgements

We thank EIKEN foundation of Japan for sharing the dataset for our research. This study was supported by JSPS KAKENHI Grant Number 19K00903.

References

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45,* 5–35.

Breiman, L. (2001). Random forests. *Machine Learning, 24,*123–140.

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA.* Amsterdam: John Benjamins.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA.

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*(4), 513–535. doi:10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using Fine-Grained clausal and phrasal indices. *The Modern Language Journal, 102*(2), 333–349. doi:10.1111/modl.12468

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. doi:10.1075/ijcl.15.4.02lu

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30,* 555–578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24,* 492–518.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity.* Honolulu: Second Language Teaching & Curriculum Center, University of Hawai'i.