Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Score Prediction by SVM and its Implication for Japanese EFL Learners' Essay Evaluation

Yuichi ONO[a*], Takeshi KATO[b] & Brendan FLANAGAN[c]
[a]*Faculty of Humanities and Social Sciences, University of Tsukuba, Japan*
[b]*Master's Course in Education, University of Tsukuba, Japan*
[c]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*ono.yuichi.ga@u.tsukuba.ac.jp

Abstract: This paper discusses the future possibilities of employing SVM and Significant Word Detection for automatic essay writing for Japanese English-as-a-Foreign-Language (EFL) Learners. After reviewing the limitations of traditional frequency-based scoring using indices related to commonly assumed constructs for learners' productive performances; that is, Complexity, Accuracy, and Fluency (CAF), this paper suggests the possibility to utilize the SVM model on Significant Word Detection instead of "frequency-based" scoring of the proposed indices on the basis of the essay data of 212 Japanese EFL learners on the Criterion Test. Specifically, the data (F-measure value) shows that the proposed model distinguish more clearly the difference in proficiency between Scores 1 and 2 on the Criterion Test in terms of indices and selected words rankings.

Keywords: SVM; Significant Word Detection; Automated Essay Scoring; CAF

1. Introduction

This paper deals with the issue about automatic scoring of L2 writing ability and proposes a possibility to employ a SVM model approach to solve this issue. Needless to say, it is an essential task for every language instructor to correctly score and evaluate learners' performance and proficiency. Since 1990, a great many studies have been published to investigate the relationship between linguistic features of the produced texts and learners' proficiency level and the most common framework has been "Complexity, Accuracy and Fluency (CAF)" (Ellis & Barkhuizen, 2005; Housen, Kuiken, & Vedder, 2012). The definition of each construct is summarized in the following table (Housen, Kuiken, & Vedder, 2012, p.2)

Table 1
Definition of CAF and Their Indices

Constructs	Definition	Indices (Example)
Complexity	The ability to use a wide and varied range of sophisticated structures and vocabulary in the L2	*Mean length of Sentence, Dependent clauses per clause, Type-Token Ratio (TTR), Sophisticated word ratio, …*
Accuracy	The ability to produce target-like and error-free language	*Errors of mechanics, Error free linguistic unit ratio, …*
Fluency	The ability to produce the L2 with native-like rapidly, pausing, hesitation, or reformulation	*Frequency of linguistic units per unit time, …*

Although its convenience in the actual teaching environment has been agreed by many researchers, the problems and challenges about the issue of CAF have been also raised by a lot of studies. The problems are around the issue of the adequacy and consistency of the constructs on the basis of CAF to predict the proficiency level (Housen & Kuiken, 2009; Pallotti, 2009). Almost all of the previous studies crucially depend on the calculation of frequency of the relevant linguistic

indices, which we will call the "Frequency-based" approach. Since indices of Complexity and Fluency are both based on "frequency", it is not clear which one they are measuring. Moreover, with this frequency-based approach, there is no consideration about the "significance" of each word's appearance on the text; in other words, one appearance is counted as one frequency equally whether or not the word is significant to the proficiency prediction. These are the motivations to start a text-mining analysis on the basis of SVM mechanism.

In the following chapter, we review the result of the frequency-based approach to the proposed indices to ask whether or not they will produce distinctive constructs which are relevant to CAF. Then, we attempt a "significance" approach to by employing a SVM model (Flanagan & Hirokawa, 2016) to prove a potential usefulness to predict proficiency.

2. Insufficiencies of Frequency-Based Approach Based on CAF Constructs

2.1 Introduction

This chapter discusses whether or not the distinctive factor structure behind Complexity and Fluency indices on the basis of 212 essays written by Japanese university students. In addition, we would like to consider what is the relationship between Complexity index values and writing ability.

2.2 Procedures

The participants wrote a persuasive essay (40 minutes, around 200 words) on the Criterion®. This is a feedback support tool for writing instruction provided by Educational Testing Service (ETS). The engine of this automated scoring engine is "e-rater" (Attali & Burstein, 2005). This is one of the most popular automated essay writing tools in Japan. In this study, the dependent variable is a score of Criterion (1-6 points). The topic was chosen from Criterion® topic list. We employed a total of 54 indices which were proposed in previous studies with two published computer programs; L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010) and Lexical Complexity Analyzer (LCA; Lu, 2012). The descriptive statistics is given below in Table 2.

Table 2
Descriptive Statistics of the Essay

Score	N	Token	Type
1	56	7852	4285
2	77	14375	7688
3	74	19195	9454
4	48	15424	7095
5	6	2749	1142
Total	261	59595	29664

The procedures of the analysis are (i) Calculation of 54 indices from essays with L2SCA and LCA; (ii) Analysis of correlation coefficients among 54 indices; (iii) Verification of the factor structure by conducting factor analysis; and (iv) Identification of the relationship between Complexity index values and essay scores provided by Criterion.

2.3 Results

The correlation networks among all indices are given in Figure 2 and that after factor analysis was described in Figure 3. (Maximum likelihood method with Promax Rotation; Cumulative % of Variance = 60%; 31 indices with factor loadings less than .400 or more than 1.00 were excluded.) The result showed that three factor structures were extracted. Factor 1 is related with the construct "Fluency". Factors 2 and 3 are related to two of the construct Complexity ("Lexical Complexity" and "Lexical Variation"). This is consistent with the posit of two subcategories within Complexity

(Bulté & Housen, 2012). However, the correlation analysis showed that, while there are some indices belonging to Factor 1: "Fluency" showed "strong or medium" correlation, most other indices belonging to Factor 2: "Syntactic Complexity" and Factor 3: "Lexical Variation" showed some "weak or no" correlation. This is shown in the Table 3 below. This crucially demonstrates that the indices referring to "Complexity" does not correlate with the score of the writing proficiency test.

3. Our Proposal: SVM and Significant Word Detection

3.1 Introduction

In this chapter, we examine the prediction of the score of the test from learners' written essays on the basis of the extraction of the significant words that are calculated from learners' sentences of different scores. The system is an application and extension of the tool proposed by Flanagan and Hirokawa (2016), which analyzes the edit distance between original and corrected learner writing sentences to automatically identify errors and extract L2 criterial lexicogrammatical features from learner corpora. In the current study, two types of corpora containing two different groups like grades 1 and 2 and grades 4 and 5 are used to determine significant words determining the scores and both positive and negative words are listed up. We need to deal with a large amount of data in an automated essay writing system, we train and evaluate the prediction performance of a Support Vector Machine (SVM) classifier by analyzing a corpus constructed using the proposed method. As to the mechanism or condition to determine significance between learners of different scores, please make reference to Flanagan and Hirokawa (2016).

3.2 Results

An SVM model trained on all features was evaluated as the baseline of prediction performance. The baseline prediction performance results are shown in Table 3 for SVM models trained by analyzing all of the features in sub-feature set. Table 3 shows the result of prediction performance for SVM model trained by analyzing all of the features of the corpus. The F-measure value of D2 vs D3 is large, accounting for higher prediction. On the contrary, that of D4 vs D5 is small, indicating that the prediction is very difficult as to Scores 4-5.

Table 3
Prediction Performance for Each Data Set

Data Set	Accuracy	F-measure	Recall	Precision
D1 vs D2	0.5736	0.4448	0.9600	0.3004
D2 vs D3	0.8532	**0.8367**	0.7057	0.6374
D3 vs D4	0.5367	0.6327	1.0000	0.4639
D4 vs D5	0.5786	0.3600	0.7000	0.2567

The top 10 positive/negative features are listed in Table 4 below for each pair of groups, illustrating that the indices among groups are different, containing indices belonging to Syntactic Variation and Lexical Variation as characterizing features predicting the score. The above table showed that the pair of D2 vs D3 showed a higher level of predictability. In this pair, seven indices are regarded as promising predictor of D3 or D2 scores as shown in Table 5 below.

This is an important implication for future possible approach to automatic evaluation. The simple correlation analysis on the basis frequency of selected indices shows only weak or no relationship with the scores. However, this does not mean that the indices are useless for indices to be employed into any kind of automatic essay evaluation system. Given an adequate process to assign a proper weight to each index might lead to a solution to more reliable predictors to learners' proficiency.

Table 4
Top 10 Positive and Negative Indices in Each Pair

	D1 vs D2		D2 vs D3	
	weight	**index**	**weight**	**index**
Positive (Upper Score's Characteristic)	0.3022032	Trait Level for Word Choice	0.1615484	Trait Level for Fluency/ Organization
	0.1569837	Trait Level for Fluency/Organization	0.1594281	LEXTOKENS*
	0.1442293	ADVV***	0.1478984	LEXTYPES*
	0.1150402	RTTR***	0.1458424	WORDTYPES*
	0.1149089	CTTR***	0.1407393	VV1***
	0.1144013	SWORDTOKENS*	0.1357381	CN*
	0.1128879	WORDTYPES*	0.1353157	W*
	0.1067614	SLEXTOKENS*	0.1350678	Trait Level for Word Choice
	0.0993168	LEXTYPES*	0.1231915	CN/T**
…….	…….		…….	
	-0.0566083	ADJV***	-0.0900191	T/S**
	-0.0596087	Number of Mechanic Errors	-0.0908619	Number of Mechanic Errors
	-0.0689778	Number of Style Errors	-0.0957847	VS1***
	-0.0714113	LD***	-0.1064915	TTR***
	-0.0737294	CN/C**	-0.1125287	VS2***
	-0.0944243	TTR***	-0.1255377	NV***
	-0.0969490	NDWZ***	-0.1356627	LS1***
	-0.1005187	NV***	-0.3045206	Number of Usage Errors
	-0.1939303	Number of Usage Errors	-0.3090309	Number of Grammar Errors
Negative (Lower Score's Characteristic)	-0.2164886	Number of Grammar Errors	-0.0900191	T/S**

	D3 vs D4		D4 vs D5	
	weight	**index**	**weight**	**index**
Positive (Upper Score's Characteristic)	0.2620786	Trait Level for Word Choice	0.1045946	CN*
	0.1514993	WORDTYPES*	0.0920999	W*
	0.1507284	LEXTOKENS*	0.0866174	CN/C**
	0.1498145	LEXTYPES*	0.0775572	LEXTYPES*
	0.1404499	CN*	0.0765693	VV1***
	0.1310414	DC*	0.0744725	WORDTYPES*
	0.1296169	W*	0.0725367	LEXTOKENS*
	0.1183451	VV1***	0.0714663	CN/T**
	0.1088334	CP*	0.0678793	Trait Level for Word Choice
…….	…….		…….	
	-0.0703645	CVV1***	-0.0187571	NV***
	-0.0904621	TTR***	-0.0265480	LOGTTR***
	-0.0943306	ADJV***	-0.0277811	Number of Mechanic Errors
	-0.105863	NV***	-0.0283895	ADVV***
	-0.1198715	LS1***	-0.0295199	TTR***
	-0.1396667	Number of Mechanic Errors	-0.0311966	LS1***
	-0.1445832	VV2***	-0.0434477	LD***
	-0.1763751	Number of Style Errors	-0.0534850	Number of Style Errors
Negative (Lower Score's Characteristic)	-0.2397661	Number of Grammar Errors	-0.0555333	Number of Grammar Errors
	-0.3041103	Number of Usage Errors	-0.0903508	Number of Usage Errors

Note. *: Index belonging to Fluency; **: Index belonging to Syntactic Complexity; ***: Index belonging to Lexical Variation.

As to the text feature, Table 5 below shows the optimal word selection performance. The optimal *N* shows that a set of 1000 top positive and negative words produces optimal prediction performance. If we look at the value of F-measure value, the pair of D1 and D2 shows the highest predictability with a comparatively small number of selections of words, while that of D4 vs D5 shows the lowest predictability. The reason may be because the number of sentences that belongs to D5 is small, as shown in Table 1 above, which probably makes it difficult to produce a prediction.

Table 5
Optimal Word Selection Prediction Performance

Data Set	N	Accuracy	F-measure
D1 vs D2	1000	0.7027	**0.7811**
D2 vs D3	500	0.6258	0.6691
D3 vs D4	200	0.6145	0.5078
D4 vs D5	3000	0.8835	0.0995

As to the pair of D1 vs D2 with the highest predictability, Table 6 shows the top 20 positive and negative words affecting the scores. As the values of Frequency and Wight shows, there is no linear correspondence between the two, suggesting that the frequency-based indices do not capture the weight features to predict the proficiency scores.

Table 6
Top 20 Positive and Negative Words for D1 vs D2

Top 10 Negative Words (Score 1 Characteristics)			Top 10 Positive Words (Score 2 Characteristics)		
Weight	**Frequency**	**Word**	**Weight**	**Frequency**	**Word**
-0.4736550	150	human	0.2764717	116	sns
-0.3860630	309	many	0.2692637	146	1:the_internet
-0.3492269	44	car	0.2642158	45	computers
-0.3141937	96	smart	0.2490998	156	technologies
-0.2978925	14	iphone	0.2486749	787	a
-0.2899323	133	life	0.2450126	15	learning
-0.2617538	35	1:to_do	0.2359601	11	1:machine_learning
-0.2395420	61	nuclear	0.2112899	318	be
-0.2390680	8	baseball	0.2078097	330	they
-0.2252266	27	check	0.1964445	30	health
-0.2238245	22	2:to_solve_this	0.1913442	190	them
-0.2230636	27	write	0.1907495	31	2:i_think_we
-0.2184093	373	our	0.1907495	32	1:think_we
-0.2179564	12	oil	0.1866921	184	invention
-0.2094453	13	drive	0.1853568	235	you
-0.1957845	58	say	0.184624	22	changed
-0.1942128	42	accident	0.1841243	37	those
-0.1901981	11	1:the_car	0.1839005	82	easily
-0.1870048	110	may	0.1825917	360	this
-0.1868908	17	1:have_many	0.1742301	81	much

4. Conclusion

This paper showed that the observable indices on the essay do not constitute the traditional CAF constructs. Moreover, it was demonstrated that the indices relating with Complexity do not have a correlation with proficiency scores. On the other hand, our proposed application of Flanagan and Hirokawa's (2016) SVM model produced a prediction to a certain degree especially toward the distinction between scores 1 and 2. As was suggested, we need to collect more data especially of Score 5, highly proficient essays, since we totally lack data of this group, which crucially makes it difficult to produce a high F-measure value of this group.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jlta.org.

Bulté, B., & Housen, A. (2012). Complexity, accuracy, and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins.

———. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26. 42–65.

Flanagan, B. & Hirokawa, S. (2016) Automatic Extraction and Prediction of Word Order Errors From Language Learning SNS, *Proc. of 5th IIAI International Congress on Advanced Applied Informatics (LTLE2016)*, 292–295.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*, 461–473.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy, and fluency: Definitions, measurement, and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp.1–20). Amsterdam: John Benjamins.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, theory and algorithms*, Kluwer Academic Publishers, 2002.

Kato, T. (2017). "Complexity and fluency indices in assessing Japanese EFL learners' writing." *Data Science in Collaboration*, 1, 82–89.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496.

———. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2): 190–208.

Nishimura, Y., Abe, D., Kawaguchi, Y., & Yao, C. (2017). The relationship between complexity and fluency in L2 writing. *The 57th National Conference of the Japan Association of Language Education and Technology*, Nagoya Gakuin University. 2017. 8.

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. doi:10.1093/applin/amp044.

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.