A Comparative Analysis of the Effectiveness of Rare Tokens on Creative Expression using ramBERT

Youbin Lee, Deokgi Kim, Byung-Won On
Kunsan National University
{hanbin0694, thekey1220, bwon}@kunsan.ac.kr

Ingyu Lee
Yeungnam University
inlee3941@gmail.com

Abstract

Until now, few studies have been explored on Automated Creative Essay Scoring (ACES), in which a pre-trained model automatically labels an essay as a *creative* or a *non-creative*. Since the creativity evaluation of essays is very subjective, each evaluator often has his or her own criteria for creativity. For this reason, quantifying creativity in essays is very challenging. In this work, as one of preliminary studies in developing a novel model for ACES, we deeply investigate the correlation between creative essays and expressiveness. Specifically, we explore how rare tokens affect the evaluation of creativity for essays. For such a journey, we present five distinct methods to extract rare tokens, and conduct a comparative study on the correlation between rare tokens and creative essay evaluation results using BERT. Our experimental results showed clear correlation between rare tokens and creative essays. In all test sets, accuracies of our rare token masking-based BERT (ramBERT) model were improved over the existing BERT model up to 14%.

1 Introduction

In the era of the Fourth Industrial Revolution, new knowledge creation based on existing knowledge is becoming more important than any other field. With the advent of Artificial Intelligence (AI) technology, which is the core of this industrial revolution, AI takes charge of simple repetitive tasks and allow humans to focus more on creative activities. So that, new innovations can be achieved through collaboration between humans and AI. This knowledge creation is based on human creative thinking (Noh and Kim, 2008)

One of the representative creative thinking activities is writing. *Creative texts* are new and at the same time communicable to readers within social and cultural contexts. *Creative writing* is the act of writing creative texts. Specifically, it is defined as a writing activity in which writers express their new and original ideas so that they can be communicated appropriately and effectively in social and cultural contexts. In this sense, through creative writing, middle and high school students can develop their ability for creativity that is essential for the Fourth Industrial Revolution.

Creativity is a human mental activity, known as a complex characteristic that is difficult to explain. In addition, it is not limited to one academic field, but the nature of creativity varies slightly depending on the disciplines in question, such as linguistics, psychology, and literature. Academically, there is no clear definition of creativity yet.

Looking at the discussion so far, it has been defined differently depending on which part of creativity the researcher has focused on. For example, (Torrance, 1974) defined creativity as "the process of becoming sensitive to problems, flaws, gaps in knowledge, missing elements, and incongruities", (Sternberg and Lubart, 1999) as "generating new and useful ideas," and (Plucker et al., 2004) as "the interaction between processes and capacities that produce novel and useful outputs within a specific social context." To date, the most widely accepted definition of creativity by many researchers is "an individual's ability to create something new and appropriate."

Table 1 summarizes the creativity evaluation metric used to evaluate actual creative writing. **Creativity in writing** is largely divided into *creativity in process* and *creativity in outcome*. The former refers to the cognitive process leading to text production, and the latter refers to creativity evaluation through the resulting text. Since computational creativity mainly focuses on the latter rather than the former, we also focus on the latter in this work.

As shown in the table, the creativity in outcome is divided into creativities in *expression* and *content*. Because clearly evaluating the creativity in content is implausible with current technology, we focus now on assessing the creativity in expression while

Findings of the Association for Computational Linguistics: ACL 2023, pages 10063–10077
July 9-14, 2023

Main category	Middle category	Subdivision	Creativity evaluation index
Creativity in Writing	Creativity in Process		Fluency
			Flexibility
			Originality
			Elaboration
	Creativity in Outcome	**Creativity in Expression**	**Originality**
			Appropriateness
		Creativity in Content	Originality
			Appropriateness

Table 1: Creativity evaluation metric.

leaving the creativity in content as a topic for our future research.

The creativity in expression has two common creativity evaluation indices. One is *originality* and the other is *appropriateness*. Technically, the originality is the ability to produce new and unique expressions that are different from existing ones, and the appropriateness must include all the requirements of 'good writing', such as having to strictly follow the grammatical rules and, in any case, having to meet the requirements appropriate for the given context.

In order to quantify the appropriateness index, various Automated Essay Scoring (AES) models that take essays as input are actively developed in the field of natural language processing in recent years. We will introduce the state-of-the-art AES models in detail in Section 2. Please note that studying the appropriateness index is out of the scope of this work.

In this paper, we focus only on the originality index in the creativity in expression. Throughout this paper, the expression includes the tokens of subword, word, n-gram, and span(phrase or sentence). Furthermore, to clearly quantify the originality index, as we already mentioned, we first consider it to be new and unique tokens within the given corpus, but the meaning of "new and unique" is ambiguous. In our second thought, we define it as *rare* tokens in the corpus.

The goal of our research is to deeply investigate the correlation between rare tokens and creativity assessment. In the corpus, rare tokens can be extracted through various methods such as Byte Pair Encoding (BPE), Inverse Document Frequency (IDF), Clustering in latent semantic spaces, SpanBERT, and existing rare word dictionaries including Stanford Rare Word (Luong et al., 2013), Cambridge Rare Word (Pilehvar et al., 2018), Contextual Rare Word (Khodak et al., 2018), and Definitional Nonce (Herbelot and Baroni, 2017).

In our framework, Bidirectional Encoder Representations from Transformers (BERT) is pre-trained with Masked Language Model (MLM). Unlike the existing BERT model, rare words rather than random words are masked and predicted in our model. Then, in the fine-tuning step, the pre-trained BERT model learns with the training set of essays.

Our contributions are the followings:

- To the best of our knowledge, we are the first to deeply study the correlation between rare tokens and creativity assessment in the automated essay scoring problem.
- We present how to extract rare tokens in various approaches: BPE, IDF, Clustering in latent semantic spaces, SpanBERT, and existing rare word dictionaries such as Stanford Rare Word, Cambridge Rare Word, and Contextual Rare Word.
- We built and validated a training set including 800 creative essays with the help of linguistics experts from the Automated Student Assessment Prize (ASAP) dataset (ASAP, 2022). Our experimental results show that all accuracies of the pre-trained BERT model with rare tokens have been improved up to 14% in assessing creativity in essays compared to the existing BERT model.

2 Related Work

The AES researches have been focused on generating hand-crafted features as an input of classification or regression (Larkey, 1998; Rudner and Liang, 2002; Attali and Burstein, 2006; Yannakoudakis et al., 2011; Chen and He, 2013; Phandi et al., 2015; Cozma et al., 2018). The linguistics features such as style and grammar are used (Lu et al., 2017; Ramalingam et al., 2018; Chen and He, 2013; Phandi et al., 2015). Sometimes, we analyze the contents by Latent semantic analysis (Ratna et al., 2007; Amalia et al., 2019; Ratna et al., 2018, 2019a,b; Shehab et al., 2018; Ratna et al., 2019c, 2015; Kakkonen et al., 2005; Xu et al., 2017), by WordNet (Al Awaida et al., 2019) and word embedding vectors (Dong and Zhang, 2016), by using specific language features (Wong and Bong, 2019; Cheon et al., 2015) and Artificial Neural Networks (Nguyen and Dery, 2016; Taghipour and Ng, 2016; Liang et al., 2018). We also hybrid the style and content analysis to improve further (Ishioka and Kameda, 2006; Peng et al., 2010; Imaki et al.,

2013; Alghamdi et al., 2014; Jin and He, 2015; Al-Jouie and Azmi, 2017; Contreras et al., 2018).

With the advent of deep learning technologies, AES have improved by using the pre-trained models with large data set (Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Wang et al., 2018; Tay et al., 2018; Farag et al., 2018; Song et al., 2020; Ridley et al., 2021; Muangkammuen and Fukumoto, 2020; Mathias et al., 2020; Jin et al., 2018; Dasgupta et al., 2018). The LSTM and RNN are naturally choices for AES task and some researchers applied BERT for the task. BERT based approaches (Uto et al., 2020; Rodriguez et al., 2019; Mayfield and Black, 2020) shows an inferior performance than LSTM (Dong et al., 2017; Tay et al., 2018) in general. However, Cao et al. (2020) and Yang et al. (2020) are knowns to show a compatible performance to LSTM based systems even with BERT. Other variations are Song et al. (2020) used a transfer learning to overcome the size limitation of training data, Wu et al. (2021) applied the R-Drop to avoid the overfitting, and Wang et al. (2022) used a transfer learning with multi-scale essay representations with BERT.

On the other hands, AES system has been studied as in a novelty or creativity detection perspective. Liang et al. (2021) proposed a model to detect creative essay using a Generative Adversarial Networks on the ASAP data. Doboli et al. (2020) used a cognitive inspired approach to detect novel ideas on short text. Chikkamath et al. (2020) applied the machine learning and deep learning approaches with various embedding vectors to find a new technology on patent data. Bhattarai et al (2020) proposed a *Tsetline* machine to detect a novel text using conjunctive clauses. Nandi and Basak (2020) proposed several CNN architectures to detect novel texts. Beaty and Johnson (2021) proposed an open platform to detect creativity based on semantic distances on word embeddings. Amplayo et al. (2019) evaluated the academic research paper novelty detection using time and impact simulations. Simpson et al. (2019) proposed Bayesian approach to predict humor and metaphor score using Gaussian process preference learning. Christophe et al. (2021; 2020) proposed a framework to detect a new topic by monitoring the geometrical properties of word embeddings. Ghosal et. al. proposed relative document vector based CNN model (Ghosal et al., 2018a) and a TAP-DLIND benchmark data sets (Ghosal et al., 2018b, 2022).

Many researches have been done in detecting creative essays as mentioned in this section. However, as authors aware, this is the first attempt to use the low frequency words to detect creative essays. Since the rare words are highly correlated with creative essays, we conjecture that the proposed approach will show a promising improvement.

3 Methodology

The ultimate goal of our study is to understand how strongly a pre-trained encoder model like BERT for ACES is affected by rare tokens.

To do this, given a large-sized set of text documents, we first extract a list of rare tokens using various approaches that we will discuss in detail in Section 3.1. Then, we will explain in detail our framework for comparative analysis of ACES in Section 3.2. Finally, in Section 3.3, we will design main questions for data-driven insights we want to know about through this study.

3.1 Extraction of Rare Tokens

Since our research focuses on *creativity in expression*, we consider various types of tokens as the expression. A token t_i^* is one of subword (t_i^s), word (t_i^w), n-gram (t_i^n), and span (t_i^S) that corresponds to a phrase, clause, or even sentence. In the case of n-gram tokens, to distinguish them from spans, 2-grams are used in this study. This is, $t_i^* \in \{t_i^s, t_i^w, t_i^{n=2}, t_i^S\}$. For example, in the pre-training step for BERT, when rare tokens are masked in a sentence like "we do not want to squander privileges and our essential things", t_i^s are 'sq', '##uan', and '##der'; t_i^w is 'squander'; $t_i^{n=2}$ is 'squander privileges'; and t_i^S is 'squander privileges and our essential things'.

A set of rare tokens $\Phi_i = \{r_1, r_2, ..., r_m\}$ is created through a method $f_i \in F = \{f_1, f_2, f_3, f_4, f_5\}$ from a corpus of large text documents $C = \{d_1, d_2, ..., d_n\}$ used to pre-train the BERT model. For a comparative study, we create a total of 7 sets of rare tokens to see if there is a correlation between creative evaluation results and rare tokens. Those sets are $\Phi_{i=1\sim7}$.

The first set of rare tokens is constructed as defined in Eq. 1.

$$\Phi_1 = \{r_i | r_i = f_1(x) \wedge r_i = t_i^s\} \quad (1)$$

, where x is a word token. f_1 is one of subword-based tokenizers. There are various to-

kenizers such as Byte-Pair Encoding (BPE), Word Piece Model (WPM), Unigram, and Sentence Piece Model (SPM), but we use BPE in this work. As the number of vocabularies increases in the pre-trained model, the dimension of word embedding vectors increases or the model becomes more complex. Therefore, units are used instead of vocabularies to reduce the number of vocabularies. A unit is a group of frequently appearing characters in a corpus and refers to a word or subword. A common word such as 'makers' and 'over' is set as one unit because it appears frequently in the corpus, while 'jet' is a rare word, so it is divided into 'j' and 'et' units. For example, in a sentence like "jet makers feud over seat width with big orders at stake", the units are {j, et, makers, fe, ud, over, seat, width, with, big, orders, at, stake}.

In the initial time, function f_1 tokenizes a given corpus. For example, {('hug', 10), ('pug', 5), ('pun', 12), ('bun', 4), ('hugs', 5)}, where each parentheses has a token and its frequency. After splitting words into characters using a pre-defined dictionary such as ['b', 'g', 'h', 'n', 'p', 's', 'u'], we get the same result as {('h' 'u' 'g', 10), ('p' 'u' 'g', 5), ('p' 'u' 'n', 12), ('b' 'u' 'n', 4), ('h' 'u' 'g' 's', 5)}. In this result, the most frequently appearing character pairs are selected. For instance, the frequency of 'hu' is 15, while that of 'ug' is 20. Since 'ug' has the highest frequency, it is newly added to the dictionary – i.e., ['b', 'g', 'h', 'n', 'p', 's', 'u', 'ug']. This process is repeated until the number of times i specified by the user. If $i = 3$, the final dictionary includes units in ['b', 'g', 'h', 'n', 'p', 's', 'u', 'ug', 'un', 'hug']. Finally, we consider the top-k units with the lowest frequency as rare tokens t_i^s.

The second set of rare tokens is created as defined in Eq. 2.

$$\Phi_2 = \{r_i | r_i = f_2(x) \wedge r_i = t_i^w\} \quad (2)$$

, where function f_2 decides if x is a rare word or not. To implement f_2, we use Inverse Document Frequency (IDF) that assigns a high score to a word that appears infrequently in the corpus, assuming it is an important word. For example, proper nouns such as 'biden' and 'google' have high values and stopwords such as 'in' and 'the' have low scores. For example, suppose that the number of documents in a corpus is one million and that of documents that contain 'biden' is one thousand. The IDF value of 'biden' $f_2(\text{'biden'}) = 1 + log(\frac{1{,}000{,}000}{1{,}000}) = 4$. As a final, we select the top-k words with the highest IDF values to use rare words for pre-training rare word-based masked language model in BERT.

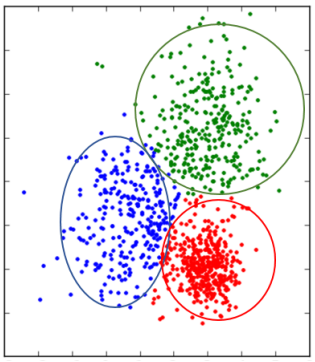

Figure 1: Overall concept of clustering contextualized vectors in the semantic space.

The third set of rare tokens is created as defined in Eq. 3.

$$\Phi_3 = \{r_i | r_i = f_3(x) \wedge r_i = t_i^w\} \quad (3)$$

, where function f_3 performs two steps. In the first step, the contextualized vector v corresponding to each word x is obtained using the pre-trained BERT and is projected into the latent semantic space. This is, $f_{BERT}(x) = v$. In the second step, all contextualized vectors are clustered in the semantic space. Since we do not know in advance how many clusters exist in the semantic space, we must use one of unsupervised clustering methods. In this work, we use Expectation-Maximization (EM), in which clustering is performed by calculating the probability of points generated by k Gaussian mixture models. In Expectation step (E-step), compute $P(C_j|v_i) = \frac{P(C_j)P(v_i|C_j)}{\Sigma_{l=1}^{k} P(C_l)P(v_i|C_l)}$. In Maximization step (M-step), for every cluster (e.g., a cluster C_j), update the weight, mean, and standard deviation of the cluster by $P(C_j) = \frac{1}{n}\Sigma_{i=1}^{n} P(C_j|v_i)$, $\mu_{C_j} = \frac{\Sigma_{i=1}^{n} v_i \cdot P(C_j|v_i)}{\Sigma_{i=1}^{n} P(C_j|v_i)}$, and $\sigma_{C_j} = \frac{\Sigma_{i=1}^{n} (v_i - \mu_{C_j})^2 P(C_j|v_i)}{\Sigma_{i=1}^{n} P(C_j|v_i)}$, using $P(C_j|v_i)$ updated in E-step.

Figure 1 illustrates the output of f_3. There exist three clusters of contextualized vectors. For convenience, we denote the clusters as green, blue, and red clusters. The red cluster has a relatively small cluster size compared to the green and blue clusters. This means that the word vectors in the green

and blue clusters are related to common topics and expressions. Words corresponding to such vectors are likely to appear frequently in a corpus. On the other hand, words belonging to the red cluster are relatively likely to be rare words in the corpus. Therefore, to extract rare words through f_3, we focus on the smallest cluster C_s (the red cluster in Figure 1). Finally, we select only words corresponding to the contextualized vectors belong to C_s. Assuming that there are three clusters C_1, C_2, and C_3, where $C_s = C_1$, all selected words must satisfy Eq. 4.

$$\{v|v \in C_s \wedge v \notin (C_2 \cup C_3)\} \quad (4)$$

The fourth set Φ_4 is the union of sets Φ_2 and Φ_3. In general, rare tokens in Φ_2 are extracted in terms of lexical representation, while those from Φ_3 are extracted in terms of semantic representation. Therefore, if BERT is pre-trained by rare word-based mask language model using Φ_4, we can know how rare words obtained by both representation approaches affect the creativity evaluation of essays. In addition, the fifth set Φ_5 is similar to Φ_3 except using n-grams instead of words. We use 2 for n in our experiments. For instance, in a sentence like "we do not want to squander", in the first step of f_3, $f_{BERT}(x) = v$, where x = 'we do', 'do not', 'not want', 'want to', or 'to squander' and the second step is the same. See Eq. 5.

$$\Phi_5 = \{r_i|r_i = f_3(x) \wedge r_i = t_i^{n=2}\} \quad (5)$$

We create the sixth set Φ_6 through function f_4 as shown in Eq. 6.

$$\Phi_6 = \{r_i|r_i = f_4(x) \wedge r_i = t_i^S\} \quad (6)$$

, where x is a sequence of words that is the input of f_4. Unlike the functions f_1, f_2, and f_3, it identifies whether a span of text is rare in a corpus. As we already discussed, creative expressions can be clauses, phrases, and even sentences, as well as subwords or words. As x is a sequence of words, we attempt to find a span of x that corresponds to a clause, a phrase, or a sentence. To present f_4, we first use SpanBERT to represent and predict spans of text, training span boundary representations to correctly predict masked span. The final loss function of SpanBERT $L(x_i)$ is to sum the losses from both Span Boundary Objective (SBO) and Masked Language Model (MLM) Objective for each token x_i in the masked span $(x_s, ..., x_e)$, where x_s and x_e are the boundary tokens. $L(x_i) = L_{MLM}(x_i) + L_{SBO}(x_i) = -logP(x_i|x_i) - logP(x_i|x_s, x_e, p_{|p(x_i)-p(x_s)|})$, where $p_{|p(x_i)-p(x_s)|}$ is the relative position between x_i's position $p(x_i)$ and x_i's position $p(x_s)$. See (Joshi et al., 2020) for details. After finding spans through SpanBERT, we use f_3 in order to detect rare spans in the corpus.

The last set of rare tokens is Φ_7, as defined in Eq. 7.

$$\Phi_7 = \{r_i|r_i = f_5(x) \wedge r_i = t_i^w\} \quad (7)$$

Recently, several dictionaries such as Stanford Rare Word, Cambridge Rare Word, Contextual Rare Word, and Definitional Nonce have been open in public. The goal of constructing such dictionaries is to get good embedding vectors from a given corpus, using Word2Vec. The most drawback of existing word embedding models is that frequent words in the corpus can generate good embedding vectors, but not for rare and unseen words. To address this problem, advanced word2vec models such as Morphological Recursive Neural Network (morphoRNN) and Neural Language Model (NLM) (Luong et al., 2013), a linear transformation of additive model $v_w^{additive} = \frac{1}{|\Gamma_w|}\Sigma_{\gamma \in \Gamma_w}\frac{1}{|\gamma|}\Sigma_{w \in \gamma} v_w$, where Γ_w is a context that contains word w, have been presented in NLP.

The function f_5 is to use one of rare word dictionaries. If x matches a rare word in a dictionary, x is added to Φ_7. For our experiment, we use the Harvard dictionary about rare words collected by Context-sensitive morphoRNN, which is the most representative one in this area.

3.2 ramBERT

In this section, we present our framework called ramBERT for comparative analysis of ACES. Figure 2 depicts the ramBERT model. First, we modify the masked language model of the existing BERT model in which rare tokens extracted through $f_i \in F$ discussed in Section 3.1 are masked and predicted in the pre-training step. Unlike existing BERT models, through the language model of ramBERT that masks and predicts rare tokens correctly, it is likely to attend more over rare tokens than over common tokens in texts. Next, the pre-trained BERT model is trained with a training set of essays in the fine-tuning step. In the final step, it automatically classifies each essay in a test

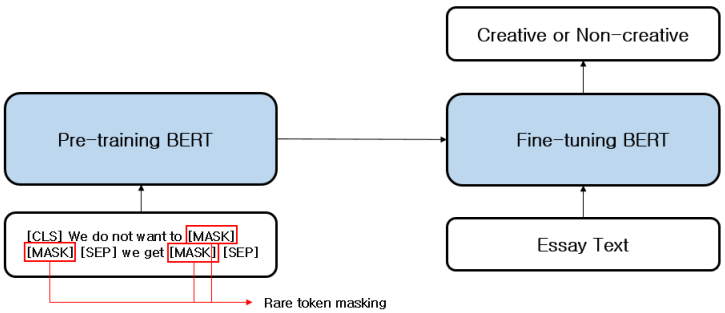

Figure 2: ramBERT: Our framework for comparative analysis of Automated Creative Essay Scoring (ACES).

set to be either *creative* and *non-creative*.

3.3 Main Questions

In this work, we will investigate the correlation between rare tokens and creative essay evaluation results. If there is any correlation, we will also examine how different types of rare tokens such as subwords, words, n-gram tokens, and spans affect creative essay evaluation. In addition, we plan to look into which type of rare token has the greatest impact on creativity assessment. Furthermore, we will find out which of $F = \{f_1, f_2, f_3, f_4, f_5\}$ for extracting rare tokens is the most effective in evaluating essays for creativity. Please, note that we now have seven sets of rare tokens $\Phi_{i=1\sim7}$. We will measure the degree of overlap of rare tokens between two sets in order to see how similar they are to each other. We will also investigate how creativity evaluation results change as the percentage of masked rare tokens is gradually increased.

4 Experimental Set-up

For the experiment, we first implemented the five rare token extraction methods. We wrote Python script codes to implement f_1 using BERTokenizer of Hugging Face, f_2 and f_3 using scikit-learn 1.2.0, f_4 using SpanBERT base model (uncased) in PyTorch, and f_5 using Stanford, Cambridge, and Contextual Rare Word dictionaries. We also modified TensorFlow code of BERT base model (uncased) for implementing the rare token-based masked language model.

A total of 4,079,432 documents from Wikipedia were collected and text was extracted by removing html tags in each document. Such refined text was used as input for BERT's pre-training. To train ramBERT, we used the same default parameters as the BERT base model, where we set 32 to batch size, 10 to epoch, 2e-5 to learning rate, and 0.1 to dropout rate. In addition, we used Adam optimizer and we set 128 to maximum sequence length, 20 to maximum number of predictions per sequence, and 0.1 to masked language mode probability.

To fine-tune ramBERT and perform the downstream task, we constructed a training set for creative essay assessment. First, we selected 800 essays at random from Prompt 1 of the ASAP dataset (ASAP, 2022). The topic of the essays is how computers affect people. In the existing ASAP dataset, the essay score ranges from 2 to 12 points, and the higher the essay score, the better the writing, regardless of creativity. Then, each essay was labelled as creative or non-creative by three domain experts who voted to classify each essay as either of creative or non-creative labels.

All models were in standalone executed in a high-performance workstation server with Intel Xeon Scalable Silver 4414 2.20GHz CPU with 40 cores, 24GB RAM, 1TB SSD, and GEFORCE RTX 3080 Ti D6 11GB BLOWER with 4,352 CUDA cores, 12GB RAM, and 7GBPS memory clock.

5 Comparative Analysis

5.1 Correlation between Rare Tokens and Creative Essay Evaluation Results

In our study, the main goal is to see if there is any correlation between rare tokens and creative essay evaluation results. Specifically, we present five rare token extraction methods $F \in \{f_1, f_2, f_3, f_4, f_5\}$. f_1 finds rare tokens based on BPE; f_2 on IDF; f_3 on clustering contextualized vectors in the semantic space; f_4 on SpanBERT; and f_5 on existing dictionaries about rare words (e.g., Stanford Rare Word dictionary).

Figure 3 shows the average accuracies of ramBERT using F. In the figure, the baseline method is the existing BERT model in which random words are masked and predicted in the pre-training step.

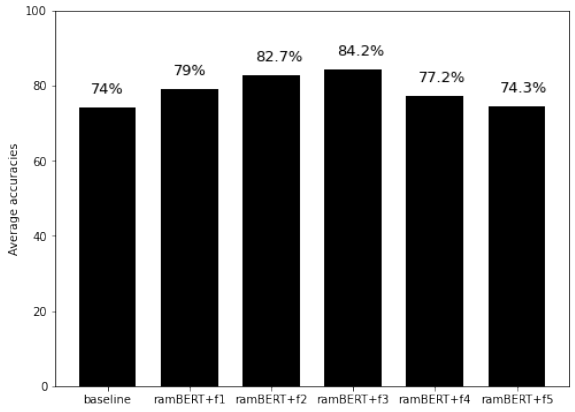

Figure 3: Average accuracies of five rare token extraction methods when the top-30% of rare tokens are masked.

	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	Φ_7
Φ_1	100.0	70.5	47.7	69.8	43.1	8.1	9.8
Φ_2	61.6	100.0	49.9	87.2	45.4	8.4	9.2
Φ_3	47.8	57.2	100.0	98.9	79.1	8.4	7.5
Φ_4	46.9	67.0	64.5	100.0	53.2	8.4	7.7
Φ_5	22.9	27.5	41.9	42.0	100.0	6.5	3.5
Φ_6	5.1	6.0	5.3	7.9	7.9	100.0	1.0
Φ_7	38.6	41.4	29.5	45.2	26.1	6.3	100.0

Table 2: ROUGE-1 of rare token sets.

	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	Φ_7
Φ_1	100.0	46.6	20.0	36.6	0.05	2.2	1.0
Φ_2	40.7	100.0	22.0	67.4	0.04	2.4	1.2
Φ_3	20.0	25.3	100.0	58.8	0.09	2.3	0.7
Φ_4	24.6	51.9	39.4	100.0	0.03	2.6	0.9
Φ_5	0.02	0.02	0.04	0.02	100.0	0.0	0.0
Φ_6	1.3	1.6	1.4	2.4	0.0	100.0	0.05
Φ_7	3.9	5.1	2.7	5.2	0.0	0.4	100.0

Table 3: ROUGE-2 of rare token sets.

The average accuracy of the baseline model is 74% or so. On the other hand, the average accuracy of using f_1, f_2, f_3, f_4, and f_5 is 79.2%, 82.7%, 84.2%, 77.2%, and 74.4%, respectively. Compared to the baseline model, ramBERT using f_1, f_2, f_3, f_4, and f_5 improved the accuracy by about 7%, 12%, 14%, 4%, and 0.5%, respectively. Surprisingly, in all cases, ramBERT significantly improved the accuracy over BERT. These experimental results indicate that even rare tokens extracted by any method $f_i \in F$ have a strong influence on creative essay assessment.

Especially, among the five methods, rare tokens extracted by f_3 seem to correlate strongly with creativity evaluation results. Please, note that ramBERT using f_3 improved the accuracy by up to 14%. Unlike subword-based tokenizing (f_1), lexical representation (f_2), and advanced word embedding (f_5) approaches, f_3 is based on clustering contextualized vectors in the semantic space. This is, rare tokens extracted through this semantic representation approach are more useful than those extracted by other extraction methods. This suggests that rare tokens extracted by considering the context of the corpus are the dominant factor in evaluating creative essays than those extracted using superficial methods such as the lexical representation approach. Furthermore, the semantic representation approach is better than the advanced word embedding method such as Context-sensitive Morphological RNN that is limited to extract rare words by affixes and frequencies.

Moreover, our hypothesis in designing f_3 is that only tokens corresponding to contextualized vectors belonging to the smallest cluster among several ones are considered to be rare in a corpus. Our experimental results showed that this hypothesis can be used to extract rare tokens that are helpful in evaluating creative essays. Consequently, Eq. 4 has been experimentally shown to be valid.

As shown in Figure 3, we can observe that rare tokens, regardless of their form, such as subwords, words, n-gram tokens, and spans, have a great impact on creativity evaluation results. However, among subwords, words, n-gram tokens ($n = 2$ in our experiments), and spans, the word-based rare tokens have the greatest impact on creative essay evaluation. The average accuracies of ramBERT with subwords (Φ_1), words (Φ_2/Φ_3/Φ_4/Φ_7), n-gram tokens (Φ_5), and spans (Φ_6) are 79%, 82.7%/84.2%/83.1%/74.3%, 79.6%, and 77.2%, when the top-30% of rare tokens are masked. Interestingly, the accuracy of ramBERT using Φ_4, the union set of rare tokens extracted by both the lexical (f_2) and semantic representation (f_3) approaches, dropped by about 1.1%. This indicates that using the semantic representation approach alone is more effective than combining lexical and semantic approaches. Another interesting point is that using word-based rare tokens improved ramBERT's accuracy rather than those in the form of n-gram tokens and spans. From these experimental results, we make sure that word-based tokens are more effective than other types of rare tokens because they are an important primitive for context understanding. For lack of space, we will discuss experimental results in more detail in Appendix A.

5.2 Characteristics of Extracted Rare Tokens

As some examples of rare tokens, Φ_1 = {'##agen', '##icult', 'dar'}, Φ_2 = {'determination', 'quar-

	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	Φ_7
Φ_1	100.0	70.5	47.7	69.7	41.8	8.0	9.8
Φ_2	61.5	100.0	49.9	87.2	44.4	8.4	9.1
Φ_3	47.7	57.2	100.0	98.9	77.9	8.4	7.4
Φ_4	46.8	67.0	64.5	100.0	52.3	8.4	7.7
Φ_5	22.2	26.9	41.3	41.3	100.0	5.3	3.3
Φ_6	5.0	6.0	5.3	7.9	6.4	100.0	1.0
Φ_7	38.6	41.0	29.3	44.9	25.0	6.3	100.0

Table 4: ROUGE-L of rare token sets.

	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6	Φ_7
Φ_1	100.0	73.7	58.3	52.2	0.0	23.6	12.3
Φ_2	74.3	100.0	57.2	74.1	0.0	29.4	12.1
Φ_3	58.3	56.6	100.0	61.2	0.0	23.0	9.9
Φ_4	57.2	76.9	67.0	100.0	0.0	38.9	10.4
Φ_5	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Φ_6	26.0	30.6	25.3	38.9	0.0	100.0	6.9
Φ_7	2.8	1.8	2.2	0.5	0.0	0.3	100.0

Table 5: BLEU of rare token sets.

antine', 'forfeiture'}, Φ_3 = {'##ian', '##vis', 'presumably'}, Φ_4 = {'##ian', 'presumably', '##gu'}, Φ_5 = {'accur_confidence', 'prefer_##uck', 'phen_##uv'}, and Φ_6 = {'prepared memorandum found in', 'engage in genuine consultations', 'pestilential burning wind called by'}, and Φ_7 = {'untracked', 'apocalyptical', 'confinement'}. Unlike other tokens, most tokens in Φ_7 seems to be extremely rare tokens across corpora. All rare tokens, regardless of their form, such as subwords, words, n-grams, and spans, are tokenized into subwords that are masked in ramBERT. For example, token 'prodigious' extracted by f_3 is tokenized into three subwords 'pro', '##dig', and '##ious' which are masked for pre-training ramBERT.

Tables 2 $\sim$ 5 show similarity values for pairs of sets. To measure the similarities, we used ROUGE-1/2/L as recall-based measure and BLEU as precision-based measure. Since both ROUGE and BLEU are the unsymmetrical metrics, the results of sim(Φ_i, Φ_j) and sim(Φ_j, Φ_i) are slightly different. In the table, since Φ_5 is a set of n-gram tokens (2-gram in our experiments), where two consecutive tokens are represented as one token, the similarity values are close to zero. Unexpectedly, we observed that rare tokens extracted by f_3 do not overlap much with those by f_2 which are similar to those by f_1. This means that rare tokens extracted by the lexical representation approach are quite different from those by the semantic representation approach.

	15%	**30%**	**50%**
f_1	77.5	79.2	80.1
f_2	80.2	82.7	83.5
f_3	81.4	84.2	85.0
f_4	75.1	77.2	77.9
f_5 (T-Rare)	74.1	74.4	75.1

Table 6: Accuracies of ramBERT according to different percentage of masked rare tokens.

Table 6 summarizes the accuracies of ramBERT according to different percentage of masked rare tokens. As the percentage of masked rare tokens increases, the accuracy of ramBERT improves, and the accuracy of ramBERT converges when the percentage of masked rare tokens is almost 50%. In f_5, we used three different dictionaries about rare words: (1) Harvard Rare Word (H-Rare), (2) Cambridge Rare Word (C-Rare), and (3) Contextual Rare Word (T-Rare). When the top-15%/30%/50% of rare words are masked, the average accuracy of f_5 using H-Rare, C-Rare, and T-Rare is 74%/74%/74.1%, 74.3%/74%/74.4%, and 75%/74.2%/75.1%, respectively. These results indicate that there is no significant difference in accuracy when using the three dictionaries.

6 Conclusion

In this paper, we proposed a method to detect creative essay writings by using a ramBERT (i.e. rare token masking-based BERT). We used seven different rare token sets and pre-trained a BERT after masking with the rare tokens on a large data. Our preliminary experimental results show that rare tokens are highly correlated with the creativity essay scores. Consequently, the ramBERT improved the accuracy up to 14% compared to a regular BERT which is based on random word masking. The performance improvements are also shown in ROGUE and BLUE scores.

Limitations

We used the ASAP data set to evaluate the performance of the proposed method. Although the dataset is well known and widely used, it has two major limitations. At first, the data size is small. Even with pre-training the model with a decently large data set (e.g., Wikipedia), the interpretation of experimental results are limited by the data size. The second limitation is an inherited bias in the data set. Since the ASAP data set is labeled by human raters, the data set is biased by personal preferences. At last, the proposed approach requires a reasonably large pre-processing to extract all the additional features which hinders a scalability. Additionally, our work is limited to only measure creativity in expression but not in content.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by Korean Government through the Ministry of Science and ICT (MSIT) under Grant NRF-2022R1A2C1011404.

References

Saeda A Al Awaida, Bassam Al-Shargabi, and Thamer Al-Rousan. 2019. Automated arabic essay grading system based on f-score and arabic worldnet. *Jordanian Journal of Computers and Information Technology*, 5(3).

Maram F Al-Jouie and Aqil M Azmi. 2017. Automated evaluation of school children essays in arabic. *Procedia Computer Science*, 117:19–22.

Mansour Alghamdi, Mohamed Alkanhal, Mohamed Al-Badrashiny, Abdulaziz Al-Qabbany, Ali Areshey, and Abdulaziz Alharbi. 2014. A hybrid automatic scoring system for arabic essays. *Ai Communications*, 27(2):103–111.

A Amalia, D Gunawan, Y Fithri, and I Aulia. 2019. Automated bahasa indonesia essay evaluation with latent semantic analysis. In *Journal of Physics: Conference Series*, volume 1235, page 012100. IOP Publishing.

Reinald Kim Amplayo, Seung-won Hwang, and Min Song. 2019. Evaluating research novelty detection: Counterfactual approaches. In *Proceedings of the thirteenth workshop on graph-based methods for natural language processing (TextGraphs-13)*, pages 124–133.

ASAP. 2022. Automated student assessment prize (asap).

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior research methods*, 53(2):757–780.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2020. Measuring the novelty of natural language text using the conjunctive clauses of a tsetlin machine text classifier. *arXiv preprint arXiv:2011.08755*.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1011–1020.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752.

Minah Cheon, Hyeong-Won Seo, Jae-Hoon Kim, Eun-Hee Noh, Kyung-Hee Sung, and EunYong Lim. 2015. An automated scoring tool for korean supply-type items based on semi-supervised learning. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 59–63.

Renukswamy Chikkamath, Markus Endres, Lavanya Bayyapu, and Christoph Hewel. 2020. An empirical study on patent novelty detection: A novel approach using machine learning and natural language processing. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–7. IEEE.

Clément Christophe, Julien Velcin, Jairo Cugliari, Manel Boumghar, and Philippe Suignard. 2021. Monitoring geometrical properties of word embeddings for detecting the emergence of new topics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 994–1003, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clément Christophe, Julien Velcin, Jairo Cugliari, Philippe Suignard, and Manel Boumghar. 2020. How to detect novelty in textual data streams? a comparative study of existing methods. In *International Workshop on Advanced Analysis and Learning on Temporal Data*, pages 110–125. Springer.

Jennifer O Contreras, Shadi Hilles, and Zainab Binti Abubakar. 2018. Automated essay scoring with ontology based on text mining and nltk tools. In *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pages 1–6. IEEE.

Mădălina Cozma, Andrei M Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102.

Simona Doboli, Jared Kenworthy, Paul Paulus, Ali Minai, and Alex Doboli. 2020. A cognitive inspired method for assessing novelty of short-text ideas. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1072–1077.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. *arXiv preprint arXiv:1804.06898*.

Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. 2018a. Novelty goes deep. a deep neural solution to document level novelty detection. In *Proceedings of the 27th international conference on Computational Linguistics*, pages 2802–2813.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty detection: A perspective from natural language processing. *Computational Linguistics*, 48(1):77–117.

Tirthankar Ghosal, Amitra Salam, Swati Tiwari, Asif Ekbal, and Pushpak Bhattacharyya. 2018b. Tap-dlnd 1.0: A corpus for document level novelty detection. *arXiv preprint arXiv:1802.06950*.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309, Copenhagen, Denmark. Association for Computational Linguistics.

Jun Imaki, Shunichi Ishihara, et al. 2013. Experimenting with a japanese automated essay scoring system in the l2 japanese environment. *Papers in Language Testing and Assessment*, 2(2):28–47.

Tsunenori Ishioka and Masayuki Kameda. 2006. Automated japanese essay scoring system based on articles written by experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 233–240.

Cancan Jin and Ben He. 2015. Utilizing latent semantic word representations for automated essay scoring. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pages 1101–1108. IEEE.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Tuomo Kakkonen, Niko Myller, Jari Timonen, and Erkki Sutinen. 2005. Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 29–36.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia. Association for Computational Linguistics.

Leah S Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95.

Guoxi Liang, Byung-Won On, Dongwon Jeong, Ali Asghar Heidari, Hyun-Chul Kim, Gyu Sang Choi, Yongchuan Shi, Qinghua Chen, and Huiling Chen. 2021. A text gan framework for creative essay recommendation. *Knowledge-Based Systems*, 232:107501.

Guoxi Liang, Byung-Won On, Dongwon Jeong, Hyun-Chul Kim, and Gyu Sang Choi. 2018. Automated essay scoring: A siamese bidirectional lstm neural network architecture. *Symmetry*, 10(12):682.

Yu-Ju Lu, Bor-Chen Kuo, and Kai-Chih Pai. 2017. Developing chinese automated essay scoring model to assess college students' essay quality. In *EDM*.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

Sandeep Mathias, Rudra Murthy, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2020. Happy are those who grade without seeing: A multi-task learning approach to grade essays using gaze behaviour. *arXiv preprint arXiv:2005.12078*.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

Panitan Muangkammuen and Fumiyo Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language*

Processing: Student Research Workshop, pages 116–123.

Dipannyta Nandi and Rohini Basak. 2020. A quest to detect novelty using deep neural nets. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Huyen Nguyen and Lucio Dery. 2016. Neural networks for automated essay grading. *CS224d Stanford Reports*, pages 1–11.

Myeong-Wan Noh and Rayeon Kim. 2008. A study on analysis of leaner responses in web board-based reading discussions. *Journal of Korea Reading Association*, 20(1):171–199.

Xingyuan Peng, Dengfeng Ke, Zhenbiao Chen, and Bo Xu. 2010. Automated chinese essay scoring using vector space models. In *2010 4th International Universal Communication Symposium*, pages 149–153. IEEE.

Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1391–1401, Brussels, Belgium. Association for Computational Linguistics.

J. A. Plucker, R. A. Beghetto, and G. T. Dow. 2004. Why isn't creativity more important to educational psychologists? potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, 39(2):83–96.

VV Ramalingam, A Pandian, Prateek Chetry, and Himanshu Nigam. 2018. Automated essay grading using machine learning algorithm. In *Journal of Physics: Conference Series*, volume 1000, page 012030. IOP Publishing.

Anak Agung Putri Ratna, Adam Arsy Arbani, Ihsan Ibrahim, F Astha Ekadiyanto, Kristofer Jehezkiel Bangun, and Prima Dewi Purnamasari. 2018. Automatic essay grading system based on latent semantic analysis with learning vector quantization and word similarity enhancement. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality*, pages 120–126.

Anak Agung Putri Ratna, Bagio Budiardjo, and Djoko Hartanto. 2007. Simple: System automatic essay assessment for indonesian language subject examination. *Makara Journal of Technology*, 11(1):2.

Anak Agung Putri Ratna, Aaliyah Kaltsum, Lea Santiar, Hanifah Khairunissa, Ihsan Ibrahim, and Prima Dewi Purnamasari. 2019a. Term frequency-inverse document frequency answer categorization with support vector machine on automatic short essay grading system with latent semantic analysis for japanese language. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 293–298. IEEE.

Anak Agung Putri Ratna, Hanifah Khairunissa, Aaliyah Kaltsum, Ihsan Ibrahim, and Prima Dewi Purnamasari. 2019b. Automatic essay grading for bahasa indonesia with support vector machine and latent semantic analysis. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 363–367. IEEE.

Anak Agung Putri Ratna, Prima Dewi Purnamasari, and Boma Anantasatya Adhi. 2015. Simple-o, the essay grading system for indonesian language using lsa method with multi-level keywords. In *The Asian Conference on Society, Education & Technology*, pages 155–164.

Anak Agung Putri Ratna, Lea Santiar, Ihsan Ibrahim, Prima Dewi Purnamasari, Dyah Lalita Luhurkinanti, and Adisa Larasati. 2019c. Latent semantic analysis and winnowing algorithm based automatic japanese short essay answer grading system comparative performance. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–7. IEEE.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Lawrence M Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Abdulaziz Shehab, Mahmoud Faroun, and Magdi Rashad. 2018. An automatic arabic essay grading system based on text similarity algorithms. *International Journal of Advanced Computer Science and Applications*, 9(3).

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728.

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. 2020. Multi-stage pre-training for automated chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 6723–6733.

R. J. Sternberg and T. I. Lubart. 1999. *The Concept of Creativity: Prospects and Paradigms*. Cambridge University Press.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

E. P. Torrance. 1974. *The Torrance Tests of Creative Thinking: Norms-technical Manual Princeton*. NJ: Personal Press.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuan-Jing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 791–797.

Wee Sian Wong and Chih How Bong. 2019. A study for the development of automated essay scoring (aes) in malaysian english test environment. *International Journal of Innovative Computing*, 9(1).

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Yanyan Xu, Dengfeng Ke, and Kaile Su. 2017. Contextualized latent semantic indexing: A new approach to automated chinese essay scoring. *Journal of Intelligent Systems*, 26(2):263–285.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

A Appendix

Table 7 shows one of high-school student's essays in the ASAP dataset. Interestingly, the baseline model classified it to be non-creative. However, no matter which rare token extraction methods are used, the ramBERT model classified it to be creative. Each token in d_i is counted as 1 if it is included in the set of tokens generated by each rare token extraction method.

The number of tokens in d_i matched to sets Φ_1, Φ_2, Φ_3, Φ_6, and Φ_7 are 6, 7, 11, 4, and 0, respectively. The tokens matched to Φ_1 are 'contr', 'distract', 'gorgeous', 'wr', '##fi', and 'distract', where the token like 'distract' appears twice in d_i. Those matched to Φ_2 are 'controversial', 'distraction', 'exposer', 'unhealthy', 'gorgeous', 'beneficial', and 'tempting'. Those matched to Φ_3 are 'contr', 'concern', 'concern', 'expose', 'gorgeous', 'concern', 'bene', '##fi', '##st', 'tempt', '##itely', and '##uter'. Those matched to Φ_6 are 'controversial issue in my', 'accessing anything', 'serious concern to', and 'tempting'.

In particular, the reason why the number of tokens matched with Φ_7 is 0 that the number of rare words in Φ_7 is as small as 2,034. Moreover, since the rare words in the Harvard dictionary were generated primarily by affixes and frequencies, it is unlikely that such rare words would appear across several domains. The examples of the Harvard dictionary is 'untracked', 'unflagging', 'unprecedented', 'apocalyptical', 'organismal', 'diagonal', 'obtainment', 'discernment', and 'confinement', where the underlined parts of the rare words are affixes.

The tokens that match Φ_2, such as 'gorgeous', 'beneficial', and 'tempting', appear to be lexically rare tokens in the corpus of essays. Most tokens matched with Φ_3, such as 'gorgeous', 'bene', '##fi', and 'tempt', are similar to them matched with Φ_2, but the number of rare tokens is relatively large. This semantic representation method (f_3) tend to extract more rare tokens in addition to the rare tokens extracted through the lexical representation method (f_2). Therefore, hidden rare tokens that could not be extracted by the existing lexical representation methods can be extracted through the semantic representation method like f_3.

From the experimental results, we can carefully hypothesize that an essay might be creative expressively if there are many rare tokens in it. Since a detailed discussion of this hypothesis is beyond the

scope of this paper, we do not proceed further here. Instead, we will deeply investigate the validity of this hypothesis through additional in-depth studies. In addition, we will attempt to establish a theory for the hypothesis.

Similarly, we observed in our experimental results that there is a correlation between essay scores and creative essays. The essay scores are evaluators' scores for how well writing is written, regardless of creativity. In the ASAP dataset, the essay scores range from 2 to 12 points, and the higher the score, the better the essay. The essay score of the essay shown in Table 7 is 12 points as well. This is because evaluators give proper scores to student essays in terms of grammar, expressiveness, and composition of writing, but in addition to them, if there is novelty in expression or content, the essay tends to be given a higher score.

I think we can all agree that computer usage is a very controversial issue in my opinion. I believe that computers have a negative effect on people. For instance, it's not safe and children can get into all sorts of things on the Internet. Also, people spend too much time in front the computer now a days it's a major distraction and also a negative effect on kids school work. It's now or never do we decide that computers have a negative effect. You decide isn't every parents biggest concern the safety of their children. When on the Internet kids are capable of accessing anything and everything. Sometimes kids don't even look for bad things they just pop up. Would you want your child viewing things that you have no control over. Also, websites like com one of the greatest concerns when it comes to Internet safety. Although you are supposed to be at least to have a most kids lie about their age. Did you know that out of users lie about their age. And it's not always a year old saying they are it could be a year old saying they're. Not only do people lie about their age they lie about who they are. Is this the kind of Internet exposer you want for you children put a stop to this right now. More than of are overweight and unhealthy. This is another negative effect computers have on people. It's a gorgeous day Bright blue skies cotton candy clouds the sun is shining and there's a nice warm breeze Perfect day to go out and get active right Wrong. None people would be inside on the computer instead of going for a walk people would spend hours on Facebook. This is a serious concern to our health. People don't exercise enough as it is and then when you add computers, people will never get active instead of playing video games online people need to be reminded that turning off the computer and playing a fun neighbourhood game of baseball is just as fun and much more beneficial. This is just one step need to take to get a healthier lifestyle. Wouldn't you agree? Did you know that kids that spend more time on computer are more likely to do poorly in school. Surely if nothing else will convince you of the negative effects of a computer, this will than coming home and doing homework more time is spent in front of the computer. As a student, I will admit that the computer is a very tempting distraction and can easily pull a student away from their studies. You can't expect a child to make the right decision and tell their they have to go because they need to study. So you do take action now, or your child will definitely suffer. The time has come to decide. Do you believe computers have a negative effect on people. It's clear that the computer is not safe. Not to mention too much time is spent on the computer instead of being active. Most importantly, computers will negatively affect children's grades. Don't wait another minute. Let's agree and do something about this.

Table 7: A student's essay d_i.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
1

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
2 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3 4

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
4

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
4

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
4

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
4

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
4