BERT-based Automatic Text Scoring for Collaborative Learning

Haruki Fukuda
Graduate School of Integrated Science and Technology
Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu,
Shizuoka 432-8011, Japan
fukuda.haruki.16@shizuoka.ac.jp

Takashi Tsunakawa
Graduate School of Integrated Science and Technology
Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu,
Shizuoka 432-8011, Japan
tuna@inf.shizuoka.ac.jp

Jun Oshima
Graduate School of Integrated Science and Technology
Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu,
Shizuoka 432-8011, Japan
joshima@inf.shizuoka.ac.jp

Ritsuko Oshima
Graduate School of Integrated Science and Technology
Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu,
Shizuoka 432-8011, Japan
roshima@inf.shizuoka.ac.jp

Masafumi Nishida
Graduate School of Integrated Science and Technology
Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu,
Shizuoka 432-8011, Japan
nishida@inf.shizuoka.ac.jp

Masafumi Nishimura
Graduate School of Integrated Science and Technology
Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu,
Shizuoka 432-8011, Japan
nisimura@inf.shizuoka.ac.jp

Abstract—**A method called Collaboration Scenario-based Scale for Emotion Regulation (CSSER) was proposed for evaluating coordination ability in collaborative learning. The method presents cartoons of several scenarios that occur in collaborative learning to the learners, and then they are asked to fill in blank balloons with their thoughts and statements about some of the scenarios; their responses are evaluated manually based on a rubric. In this paper, we propose a BERT-based classification model for automatic text scoring in CSSER, which enables us to group learners based on their abilities measured by CSSER. Further, we demonstrate the effectiveness of data-augmentation techniques for the proposed method using lexical substitution and back-translation.**

Keywords—Automatic text scoring, BERT, Collaborative learning, Data augmentation

I. Introduction

Collaborative learning is a group-learning activity where multiple learners work together and exchange opinions. Therefore, both their socio-epistemic and socio-emotional coordination are required for effective team work. [1] Socio-epistemic coordination refers to the adjustment to socio-epistemic problems, such as discussing the solution of a given task, selecting promising ideas, and exchanging understanding of expertise in a task. In contrast, the socio-emotional coordination includes adjustment activities for the problems of human relationships that arise from conflicts between the members of a group. Socio-emotional issues occur more frequently in collaborative learning than in traditional learning environments. [2] Thus, socio-emotional coordination is essential for learners to engage in productive collaborative learning. [3] It should also be considered that coordination skills vary among different learners whereas the priority of the socio-emotional and socio-cognitive aspects depends on the learner. More specifically, in knowledge construction situations, the learning process is highly dependent on the learner's autonomy, and their coordination skills have a significant impact on learning effectiveness. Therefore, their identification and understanding is particularly important for teachers.

Collaboration scenario-based scale for emotion regulation (CSSER) has been proposed as a method for evaluating coordination ability in collaborative learning [4]. Human raters use a rubric to evaluate texts written by learners: however, this is not practical for a large number of learners as the evaluation is performed manually.

Several studies have been conducted on text scoring in education using various approaches such as syntactic/semantic parsing and machine learning. For the automatic scoring of English texts written by non-native English-language learners, Landauer et al. [5] proposed a method that uses latent semantic analysis (LSA). This method determines the most-similar text using LSA-based semantic similarities, and it assigns a score based on the most-similar text. Alikaniotis et al. [6] showed that recurrent neural-network models are effective for improving performance in similar tasks. In recent years, text scoring has commonly been treated as a supervised text-classification task [6].

Bidirectional encoder representation from transformers (BERT), which performs well on various natural-language processing tasks, exhibits high performance in supervised text classification [7]. Owing to pretraining with a large amount of text corpora, BERT can achieve relatively high performance with small amounts of task-specific data. However, it is difficult to prepare sufficient training data for a specific task such as CSSER.

Recently, several data-augmentation methods have been proposed to improve the performance and robustness in tasks without sufficient training data. In particular, Kumar *et al.* [8] have reported the effectiveness of data augmentation in some tasks using pretrained transformers such as BERT. Yu *et al.* [9] showed that back translation can generate various paraphrases while retaining the original meaning, thereby improving the performance in question-answering tasks.

In this paper, we propose a BERT-based classification model for automatic text scoring, which allows the early identification of the learners' cooperation skills, leading to more efficient group organization and learning support. Moreover, our study demonstrates the effectiveness of data-augmentation techniques using lexical substitution and back-translation for a BERT-based classification model.

This work was supported by JSPS KAKENHI Grant Number JP19H01714.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

II. Proposed Method

A. CSSER

Each of the four CSSER scenarios created by Oshima et al. [4] contains a typical scene in collaborative learning; the story is described as frames that represent conversations among group members of a particular subject. At the end of each scenario, a frame is presented with blank balloons for the character that represents the subject, and the subject is then asked to fill in what he/she thinks hereafter referred to as “voice of the mind” and what he/she says in the situation hereinafter termed the “statement.” Ratings from 1 to 5 are assigned to these texts based on two aspects of the predefined rubric: socio-epistemic and socio-emotional.

B. Text scoring by BERT-based classification

BERT is a general-purpose language-representation model based on a transformer [7], which uses self-attention mechanism. By pretraining a large text corpus and fine-tuning for each task, BERT achieves state-of-the-art performance in many natural language processing tasks.

We use a model pretrained with Japanese Wikipedia articles comprising approximately 18 million sentences. The half-width characters in the text are normalized to full-width characters and tokenized into subwords using Juman++ [10], followed by the application of byte pair encoding [11]. We then fine-tune the model so that the input is the text of the CSSER “voice of the mind” or “statement” and the output is a score to be one from 1 to 5 for both the socio-epistemic and socio-emotional aspects. The score with the highest likelihood is adopted as the result of the automatic scoring.

C. Data augmentation

1) Lexical substitution

BERT is trained on a large amount of text using a pretraining task called “masked language modeling” wherein the model has to predict masked words based on the context. [7] We use this task to generate a new sentence by masking some parts of the text and substituting them with predicted tokens for each mask using a pretrained BERT model. We mask 10% of the words chosen randomly in each sequence by replacing them with a [MASK] token. The model then attempts to predict the original masked words based on context that comprises surrounding non-masked words in the sequence.

2) Generation of back-translated paraphrases

Using the v3 NMT model of the Google Translate API, we generate paraphrases of the original text by translating it from Japanese into Korean, Chinese (simplified), and English, and then we translate the results back into Japanese. We can use the back-translated Japanese texts as paraphrase if a different text is obtained assuming the meaning of the sentence remains unchanged in the translation processes.

3) Paraphrase filtering

Augmented training data includes noise, which has a negative influence on training, to classify natural paraphrased text, we use BERT as a language model [12] to calculate the perplexity of each paraphrased text, and we use the text with the relatively smallest perplexity as the paraphrased text.

TABLE I. THE DISTRIBUTION OF ANNOTATION BY MANUAL SCORING

	Socio-emotional		**Socio-epistemic**	
	Statement	***Voice of the mind***	***Statement***	***Voice of the mind***
1	46 (0.91%)	83 (1.59%)	330 (6.52%)	407 (7.79%)
2	927 (18.29%)	510 (9.76%)	665 (13.13%)	957 (18.31%)
3	562 (11.09%)	2505 (47.93%)	1677 (33.12%)	2120 (40.57%)
4	1895 (37.40%)	1488 (28.47%)	1939 (38.29%)	1512 (28.93%)
5	1637 (32.31%)	640 (12.25%)	453 (8.95%)	230 (4.40%)
total	5067	5226	5064	5226

TABLE II. THE NUMBER OF ADDED PARAPHRASED TEXT

		Socio-emotional		**Socio-epistemic**	
		Statement	***Voice of the mind***	***Statement***	***Voice of the mind***
	Annotated data	5067	5226	5064	5226
Lexical substitution	25%	1267	1307	1266	1307
	50%	2534	2613	2532	2613
	75%	3800	3920	3798	3920
	100%	5067	5226	5064	5226
Back-translated	25%	3800	3920	3798	3920
	50%	7601	7839	7596	7839
	75%	11401	11759	11394	11759
	100%	15201	15678	15192	15678

III. Experiment

A. Setup

We used a dataset that consisted of group work; we obtained the dataset from two CSSER experiments conducted during the first lecture and the last lecture of a semester in a seminar at the Shizuoka University from 2015-2017. A total of 10293 responses were obtained for the “voice of the mind” and “statement,” and the CSSER scores for each response were annotated manually. The data collected in 2015 were rated by two researchers familiar with the coordination of learning activities based on the CSSER rubric, and the scoring mismatch was adjusted after a discussion. In contrast, the data collected in 2016 and 2017 were divided and scored by two raters, who adjusted the grading criteria using the sample data and reached a sufficient consensus. The distribution of the manually annotated scores is shown in Table I. The response data for each of these scenarios and the annotated scores were used as experimental data.

The data for the experiment were randomly decided into training, validation, and test sets in a 3:1:1: ratio. We added R [%] of the paraphrased text in the ascending order of perplexity to the training set; R was either 25%, 50%, 75%, or 100%. We conducted an experiment without using paraphrased text (i.e., R = 0%). Table II shows the number of annotated data and additional paraphrased text for each percentage. The back-translated paraphrases are created in three languages; therefore, three times more paraphrase is added as a lexical substitution.

Cohen's coefficients of agreement (κ coefficients) and micro-averages of the F values were used as evaluation measures. We set the vocabulary size (including subwords) to

32000, batch size to 32, and epoch number to 3 as the main parameters of BERT. As a baseline, we compared the proposed method with a method using Doc2Vec [13], which comprises three steps. First, all texts are converted from "voice of the mind" and "statement" into vectors by Doc2Vec. Then, the most-similar text in the training data to the input text is identified by calculating the cosine similarities between vectors. Finally, the score of the most-similar text is extracted.

B. Results

The results of the comparison of the baseline method and the BERT classifications are shown in Table III where it is clear that for all data, the BERT classifications outperformed the baseline (Doc2Vec) method.

BERT outputs the likelihood for each classification, which indicates how likely the text is classified into the class. The likelihood obtained by BERT for each of the ratings from 1 to 5 varied among the input texts, with some texts having very high likelihood for a single rating and others with close top rating values. We hypothesized that the higher likelihood leads to higher rating performance. Thus, the kappa coefficients were calculated only for the top input texts, which were ranked in descending order based on the likelihood of the ratings output in the test set. Fig. 1 shows the results obtained using the socio-emotional aspects of the "Work and Communication" data, which is one of the four scenarios. The vertical axis is the kappa coefficient whereas the horizontal axis represents the percentage of data used to calculate the kappa coefficient in descending order based on the likelihood. Therefore, it was confirmed that the higher the likelihood of the text, the higher the grading performance.

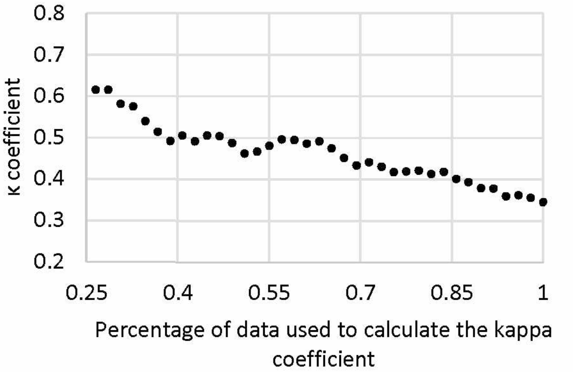

Fig. 1. Kappa coefficients calculated only for high likelihood input text

Tables IV and V list the experimental results for all four scenarios. The BERT-based classification model exhibited a performance of 0.348 on average in terms of Cohen's kappa coefficient compared with human evaluators. For both the socio-emotional and socio-epistemic aspects, the performance was slightly better when paraphrased text was included as training data.

In addition to the socio-epistemic aspects of the statement, the results showed that the performance is higher when 25%, 50%, or 75% of the training data is included compared to when the entire paraphrased text is included in the training data. Consequently, filtering paraphrases with perplexity is an effective approach.

TABLE III. TEXT SCORING RESULTS OF THE COMPARISON OF THE PROPOSED METHOD (BERT-BASED CLASSIFICATION) AND BASELINE(DOC2VEC)

		Statement		**Voice of the mind**	
		κ coefficient	***F-value***	***κ coefficient***	***F-value***
Socio-emotional	BERT	0.348	0.570	0.242	0.516
	Doc2Vec	0.162	0.387	0.113	0.356
Socio-epistemic	BERT	0.349	0.579	0.182	0.531
	Doc2Vec	0.173	0.395	0.111	0.426

TABLE IV. TEXT SCORING RESULTS WITH DATA AUGMENTATION BY USING LEXICAL SUBSTITUTION

	R	**Statement**		**Voice of the mind**	
		κ coefficient	***F-value***	***κ coefficient***	***F-value***
Socio-emotional	0%	0.348	0.570	0.242	0.516
	25%	0.398	0.599	0.266	0.532
	50%	0.436	0.612	0.260	0.530
	75%	0.314	0.545	0.239	0.560
	100%	0.313	0.543	0.219	0.524
Socio-epistemic	0%	0.349	0.579	0.182	0.531
	25%	0.372	0.588	0.215	0.533
	50%	0.373	0.590	0.211	0.520
	75%	0.330	0.561	0.214	0.522
	100%	0.336	0.565	0.204	0.520

TABLE V. TEXT SCORING RESULTS WITH DATA AUGMENTATION BY USING BACK-TRANSLATION

	R	**Statement**		**Voice of the mind**	
		κ coefficient	***F-value***	***κ coefficient***	***F-value***
Socio-emotional	0%	0.348	0.570	0.242	0.516
	25%	0.344	0.562	0.257	0.522
	50%	0.358	0.570	0.256	0.522
	75%	0.348	0.567	0.258	0.515
	100%	0.349	0.567	0.257	0.511
Socio-epistemic	0%	0.349	0.579	0.182	0.531
	25%	0.385	0.592	0.218	0.533
	50%	0.371	0.585	0.221	0.559
	75%	0.388	0.590	0.244	0.560
	100%	0.395	0.597	0.230	0.546

IV. CONCLUSIONS

We proposed a method to automate text scoring in CSSER using a BERT-based classification model that enables immediate feedback of the CSSER evaluation result. Experimental results showed that the BERT-based classification model outperformed the baseline Doc2Vec method. To compensate for the lack of training data, we introduced back-translation and lexical substitution using BERT s training data, with the paraphrased text filtered by perplexity. The experimental results show that filtering paraphrases with perplexity is effective against a lack of training data. The proposed method achieved κ = 0.405 (statement) and κ = 0.255 (voice of the mind) compared with those for human raters.

Moreover, the BERT-based classification model and data augmentation method using the paraphrased text filtered by perplexity are applicable to any other text scoring tasks that do not have sufficient annotation data.

In the future, we plan to explore various models for further performance improvement. Another future direction is applying the proposed method to aid the manual scoring of CSSER by displaying text similar to the text to be evaluated with its score in the training data.

References

[1] H. Järvenoja, S. Volet, and S. Järvelä, “Regulation of emotions in socially challenging learning situations: an instrument to measure the adaptive and social nature of the regulation process,” Educational Psychology, vol. 33, no. 1, pp. 31–58, Jan. 2013, doi: 10.1080/01443410.2012.742334.

[2] H. Järvenoja and S. Järvelä, “Emotion control in collaborative learning situations: Do students regulate emotions evoked by social challenges?,” The British journal of educational psychology, vol. 79, pp. 463–481, Mar. 2009, doi: 10.1348/000709909X402811.

[3] P. van den Bossche, M. R.Segers, and P. Kirschner, “Social and cognitive factors driving teamwork in collaborative learning environments : Team learning beliefs and behaviors,” Small Group Research, Jan. 2006.

[4] R. Oshima and J. Oshima, “Collaboration scenario-based scale for emotion regulation: Measuring learners’ agency to regulate own, others’ and group emotions,” in Proc. of EdMedia + Innovate Learning 2015, Jun. 2015, pp. 796-801, [Online]. Available: https://www.learntechlib.org/p/151349.

[5] T. K. Landauer, “Automated scoring and annotation of essays with the intelligent essay assessor,” Automated essay scoring : A cross-disciplinary perspective, pp. 87–112, Jun. 2003, [Online]. Available: https://ci.nii.ac.jp/naid/10030004395/.

[6] D. Alikaniotis, H. Yannakoudakis, and M. Rei, “Automatic text scoring using neural networks,” arXiv preprint arXiv:1606.04289, 2016.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv: 1810.04805, 2018.

[8] V. Kumar, A. Choudhary, and E. Cho, “Data augmentation using pre-trained transformer models.” arXiv preprint arXiv:2003.02245, 2020.

[9] A.W. Yu, D. Dohan, M.T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q.V. Le, “Qanet: Combining local convolution with global self-attention for reading comprehension,” arXiv preprint arXiv:1804.09541, 2018.

[10] A. Tolmachev, D. Kawahara, and S. Kurohashi, “Juman++: A morphological analysis toolkit for scriptio continua,” in EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing, 2018, pp. 54–59, doi: 10.18653/v1/d18-2010.

[11] R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Aug. 2016, pp. 1715–1725, doi: 10.18653/v1/P16-1162.

[12] X. Song, “BERT as Language Model,” GitHub repository. GitHub, 2019, [Online]. Available: https://github.com/xu-song/bert-as-language-model.

[13] Q. v. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” arXiv preprint arXiv: 1405.4053, 2014.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE