Study on Automatic Scoring of Descriptive Type Tests using Text Similarity Calculations

Izuru Nogaito
KDDI R&D Laboratories Inc.
3-10-10 Iidabashi,
Chiyoda-ku, Tokyo, 102-8460 Japan
+81-3-6678-1609
iz-nogaito@kddilabs.jp

Keiji Yasuda
KDDI R&D Laboratories Inc.
3-10-10 Iidabashi,
Chiyoda-ku, Tokyo, 102-8460 Japan
+81-3-6678-1609
ke-yasuda@kddilabs.jp

Hiroaki Kimura
KDDI R&D Laboratories Inc.
3-10-10 Iidabashi,
Chiyoda-ku, Tokyo, 102-8460 Japan
+81-3-6678-1609
ha-kimura@kddilabs.jp

ABSTRACT

In this paper, we evaluate the automatic scoring of a descriptive type test. In the experiments, three test similarity measures are compared in terms of automatic scoring quality. Two of them are BLEU and RIBES, which are *n*-gram and word-level matching processes respectively, originally used for automatic evaluation of machine translation output. The other similarity process is Doc2Vec, which utilizes distributed representation to calculate the cosine distance. It was finally found that, according to the experimental results, the most efficient process used to calculate the text similarity depends on the type of the question.

Keywords

Doc2Vec, BLEU, RIBES, Text Similarity, auto-scoring

1. INTRODUCTION

Recently, the importance of "21st Century Skills" has been advocated in educational circles. A descriptive type of test is one of the methods to measure this skill; hence, this type of test is becoming more important than a multiple choice test.

In this paper, we carried out experiments on automatic scoring of a descriptive type test. There are two types of methods for automatic descriptive type test scoring. The first method is a similarity-based method, which computes the similarity between a student's answer and a model answer. The second method does not require a model answer; however, it requires several natural language processing (NLP) tools that compute cohesion, coherence, etc. [1]. In this research, we adopt the first approach because our target language for automatic scoring is Japanese and some of the NLP tools are not supported in Japanese. Furthermore, our research partner could provide test items and model answers. In this paper, section 2 describes similarity measures that are used for automatic scoring. Section 3 demonstrates the experiments and their corresponding results, and finally, section 4 describes the conclusions and future work.

2. SIMILARITY MEASURES

In this research, we apply two similarity measures based on surface expression. Both of them were proposed for automatic evaluation of machine translation output. We also apply the similarity measures in a distributed expression to the automatic scoring experiments. In this subsection, we explain these similarity measures.

2.1 Similarity in surface expression

BLEU [2] is proposed for the evaluation of machine translations. It uses *n*-gram matching between a reference sentence and a machine translation output. A sentence that is shorter compared to the reference is penalized in the BLEU score calculation.

RIBES [3] is also an automatic evaluation measure for machine translations. First, it compares the machine translation output with a reference at the word level. Then, it inspects the word order for common words based on the rank correlation coefficient.

2.2 Similarity in distributed expression

Recently, by using deep learning technology, a word or sentence can be converted into a distributed expression that is a vector of several hundred dimensions. According to previous research [4, 5], the cosine similarity between the distributed expressions is fairly close to a semantic similarity. In this research, the gensim [1] version of Doc2Vec is used to build the model that converts the document into a distributed expression.

Table 1: Statistics of the Training Corpus for Doc2Vec

	# of words	Lexicon size
Japanese wiki abstract (WIKI)	29,944,313	1,398,558
Mainichi-News-Paper (1991-2014) (NP)	504,844,192	5,578,327
WIKI + NP	534,788,505	6,376,935

3. EXPERIMENTS

3.1 Experimental settings

Doc2Vec requires a text corpus for model training. For the experiments, we use a Wikipedia corpus (WIKI) and a Mainichi Newspaper corpus (NP). In addition, three models are trained: one using WIKI, one using NP and one using both WIKI and NP. Then, the best model is chosen for each test item in terms of the automatic scoring performance. Table 1 demonstrates the statistics of each particular corpus. In the experiments, we use ten

[1] https://radimrehurek.com/gensim/

test items.

Table 2 Answer Text-Data Specification

Item ID	Topic of question	Question type	Ave. length of student answers (words)	Lexicon size of student answers	Number of students
ID01	Book	Graph reading	112.2	62.5	21
ID02	Fisherman	Summarization	49.7	33.4	21
ID03	Food	Graph reading	96.4	49.0	24
ID04	Fishery	Graph reading	87.8	53.5	22
ID05	Supermarket	Summarization	101.4	59.7	22
ID06	University	Summarization	110.7	71.6	20
ID07	Japanese	Summarization	77.7	46.8	32
ID08	Mail	Summarization	58.9	44.6	42
ID09	Vietnam	Graph reading	57.5	31.2	29
ID10	Beef	Graph reading	90.2	44.2	24
ID01-10	Average		84.3	49.6	25.7

All test items are answered by at least twenty students, aged between 10 and 16 years. Each question has its own target grade. Table 2 demonstrates the data set. In the table, "Graph reading" indicates the situation where the students are asked to describe a fact that can be read from the given graphs. Normally this type of question is a short sentence. Further, "Summarization" indicates the situation where the students are asked to summarize a given text between 300 to 800 words long. In each test item, four model answers are made by four teachers. Each answer is also scored by four teachers. Averaged scores are used as the recorded evaluation results in the experiments.

3.2 Experimental results and Discussion

Figure1 shows the correlation between the subjective score and automatic similarity. For Doc2Vec, we trained models with three conditions: Newspaper corpus only (D2V/NP), Wikipedia corpus only (D2V/WIKI) and both Newspaper and Wikipedia (D2V/NP + WIKI).

The methods that use similarity in surface expression are partly advantageous in the summarization question type. In this type of question, students tend to use the expression in the given question sentence, and the variety of their word choice is small. Thus, the possibility of matching words on the model answer could be high. In fact, the correlation values of BLEU and RIBES for ID02, ID05, ID06, ID07 and ID08 are relatively high.

The methods that use similarity in distributed expression are partly advantageous for the automatic scoring of graph reading questions. In general, the answer for this kind of question has a wide variation of words because students are free to choose their own words.

Both types of results, however, are shown on the graph of reading questions. First, the correlation value from Doc2Vec is better than the other methods for ID03, ID04 and ID10. This is due to the reason described previously. Second, the value of Doc2Vec is inferior for ID01, though it is a graph reading question. In this case, we understand that the corpus used does not share many similar words with the model answer sentences. The result also shows that the Doc2Vec similarity sometimes also works as a complemental similarity.

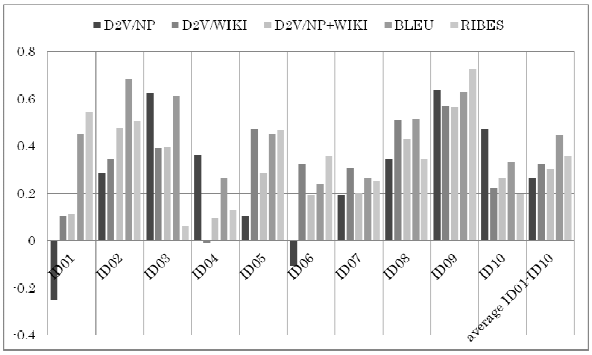

Figure 1 Correlation between subjective score and automatic method

4. CONCLUSIONS AND FUTURE WORK

For automatic scoring, we compared the Doc2Vec, the BLEU, and the RIBES similarities. In the case where the answers include a wide variation of words among students, the method using distributed expression seems to be more advantageous.

In future work, we will conduct research to use several similarities in a complementary way. We will also compare several methods, including the method using cohesion and coherence [1] that is described in the introduction section as a second method.

5. ACKNOWLEDGMENTS

This work uses model answers, student's answers, and scoring data that came from the Lojim clam school. (http://lojim.jp/).

6. REFERENCES

[1] Scott A. Crossley, Danielle S. McNamara.: Cohesion, coherence, and expert evaluations of writing proficiency, Proc. of the 32nd annual conference of the Cognitive Science Society, pp. 984-989, 2010.

[2] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, in Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL), pp. 311–318 (2002)

[3] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada.: Automatic Evaluation of Translation Quality for Distant Language Pairs, Conference on Empirical Methods on Natural Language Processing (EMNLP), Oct. 2010.

[4] Tomas Mikolov and Kai Chen and Greg Corrado and Jeffrey Dean.: Efficient Estimation of Word Representations in Vector Space, http://arxiv.org/pdf/1301.3781.pdf

[5] Quoc Le, Tomas Mikolov.: Distributed Representations of Sentences and Documents, http://arxiv.org/abs/1405.4053