So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education. Australia: Asia-Pacific Society for Computers in Education

Applying Key Concepts Extraction for Evaluating the Quality of Students' Highlights on e-Book

Albert YANG[a*], Irene Y.L. CHEN[b], Brendan FLANAGAN[c] & Hiroaki OGATA[c]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Department of Accounting, National Changhua University of Education, Taiwan*
[c]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*yang.ming.35e@st.kyoto-u.ac.jp

Abstract: The quality of students' highlights can be an indicator of their learning performance. While the most common approach to grade their highlights is by humans, human grading can be inconsistent, especially when the number of highlights are large or when graders have different background knowledge. In this research, we propose a model to automatically extract important concepts from class materials, analyze students' highlights and find the correlation between highlight quality and students' learning performance. We first compared different text summarization algorithms with different evaluations to see which of them generates the summarization that is closest to the reference answer generated by humans. Then we used the selected algorithm to summarize the text from learning materials as important concepts, and compared the summaries with students' highlights to calculate their highlight scores. Finally, we considered the highlight score from the best method as the highlight quality and observed whether it has a correlation to students' learning performance.

Keywords: E-learning, text summarization, learning analytics

1. Introduction

In this study, we analyzed students' highlight records of a 12-week course in a university. There were 44 students enrolled in this course. A total of 22 slides of learning materials were uploaded to BookRoll. BookRoll is a digital learning material (e-Book) reading system (Flanagan & Ogata, 2017) that provides students with e-Books and records their behaviors while reading. Students' e-Book reading actions in BookRoll have been described in detail by (Ogata et al., 2015). Before each class, the instructor uploaded the slides of the class to BookRoll for students to preview. Students were asked to highlight the words or sentences they think are important using the marker function on BookRoll. There are two types of markers they can use, red marker for important concepts and yellow marker for concepts they feel difficult to understand. The instructor provided a reference answer which is the key concepts highlighted by the instructor every week. We then used this reference answer to test different summarization techniques, evaluate students' highlights, and find the relationship between highlight quality and students' learning performance. This study aims to answer the following research questions:

RQ1: What is the best key concepts extraction algorithm for evaluating the quality of students' highlights on e-Book?

RQ2: Does the quality of students' highlights affect their learning performance?

2. Method

2.1 Preprocessing

Since the content of the slides uploaded by the instructor is in PDF format, it needs to be converted to

raw text for our analysis. We used python's pdfminer package to convert the content to a plain text file. Both instructor and students' highlights were collected in text form using BookRoll. We applied preprocessing techniques on the raw text of slides and highlights, such as removing special characters and converting text to lowercase. In addition, since the students may delete a marker after adding it, there are two types of records in the e-Book, ADD_MARKER and DELETE_MARKER. In order to ensure that the records we use are the highlights reserved by the students, we ignored the highlights which exist in both records.

2.2 Methods for Key Concepts Extraction

After preprocessing the content of the slides and the highlights, we extracted the important concepts from the slides using text summarization algorithms. TextRank (Mihalcea & Tarau, 2004) and RAKE (Rose et al., 2010), which belong to traditional machine learning methods, and BERT (Devlin et al., 2018), a deep learning architecture were used in this study.

For TextRank, we cut the text into multiple sentences, and selected a quarter of sentences as a threshold to represent the key concepts. TextRank can be used to extract either keywords or key sentences from text.

For RAKE, the full text was passed to the model and a quarter of phrases were extracted as the threshold to represent the key concepts.

For BERT, two approaches were used to tokenize the class materials. We tokenized the text into sentences and pages. To cut the text into pages, we concatenated every sentence in each page as a long sentence. Our approach is to tokenize the incoming text into sentences or pages, pass the tokenized sentences or pages to BERT for inference to output embeddings, and then cluster the embeddings with K-Means. Since the embeddings include more than 700 dimensions which cannot be passed directly to K-Means, we applied PCA to reduce the number of dimensions to two. Centroids of the clusters in the vector space represent the key concepts in the original text. For each key concept, we selected the embedding sentence or page that is closest to the centroid as the represented sentence or page. Figure 1 shows the result of concepts clustering for slides using K-Means.

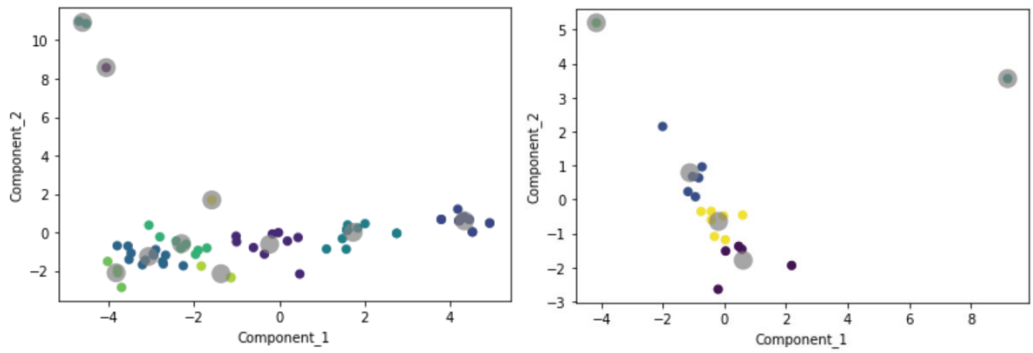

Figure 1. Using K-Means to cluster the embedding sentences (left) and pages (right). Dark centroids represent the key concepts in the slides. Each points represents a sentence (left) or a page (right). X axis and Y axis are two principle components generated by PCA.

After extracting key concepts from the learning materials, we used the highlights provided by instructor as reference answer to evaluate the quality of summaries from machine using BLEU 1, BLEU 2, BLEU 3, BLEU 4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), and ROUGE (Lin, 2004).

3. Results and Discussion

For Research question 1, What is the best key concepts extraction algorithm for evaluating the quality of students' highlights on e-Book?

We tested TextRank, RAKE, and two variants of BERT as different text summarization

techniques and compared the summaries with reference answers using standard metric BLEU, METEOR and ROUGE. The best results were obtained using the traditional machine learning method, RAKE. It outperforms other three methods in all metrics except BLEU 4 and ROUGE-L. This is expected as the summaries of RAKE are phrases while others are sentences. BLEU 4 takes into account the 4-gram co-occurrence while the length of phrases generated by RAKE is often shorter than the sentences generated by other methods. Similarly, ROUGE-L measures the longest common sequence which is also bad for RAKE. The results obtained for the different algorithms we used are showed in Table 1.

Table 1. *Evaluation results for the four text summarization techniques.*

	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE-L
TextRank	2.55	1.74	1.02	0.56	1.63	2.24
RAKE	**2.87**	**2.17**	**1.31**	0.43	**2.27**	2.29
BERT_SENTENCE	2.83	1.99	1.26	**0.67**	2.21	**2.73**
BERT_PAGE	2.84	1.96	1.18	0.54	1.57	1.89

Surprisingly, the deep learning methods, BERT, do not outperform the traditional machine learning method, RAKE. It only performs the best when the metrics evaluate longer common sequences such as BLEU 4 and ROUGE-L. The possible reason is the data being analyzed in this study. One advantage of using deep learning models is the ability to analyze semantic meanings of the sentences. However, most of the content in the slides are sentences extracted from different paragraphs in papers or textbooks, and the meaning of text in each page is very different from another, which means the contextual meaning in sentences is lacking. Therefore, it is hard to take the advantage of BERT when the learning materials consist of phrases or incomplete sentences. We therefore suggest to represent the input in full text to leverage BERT. For instance, a textbook or papers.

For Research question 2, Does the quality of students' highlights affect their learning performance?

We adopted RAKE as the text summarization method for our study and used the summaries as key concepts to evaluate the quality of students' highlights. BLEU 1 is used to calculate the similarity between summaries and highlights since most of the highlights contain words or phrases and the length is often less than 2. The sum of highlight scores of 11 weeks is regarded as the quality level of highlights. Figure 2 shows the correlation between highlight score and students' learning performance. The Spearman correlation is 0.75 with P-value less than 0.001, which indicates that the highlight score is highly correlated to learning performance. We assigned students into two groups. The top 20% students belong to HIGH_PERFORMANCE and others belong to LOW_PERFORMANCE. Since both groups show the normal distribution, a one-way ANOVA test is performed. Table 2 shows that the mean of high learning performance group is larger than low learning performance group at a statistical significance level.

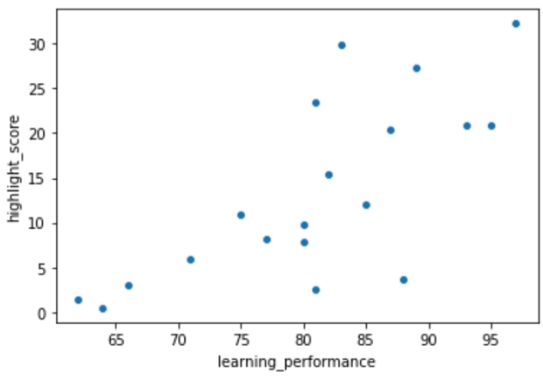

Figure 2. The correlation between highlight score and students' learning performance. The Spearman

correlation is 0.75, $p < 0.001$.

Table 2. *The one-way ANOVA test for group with high learning performance and group with low learning performance.*

	Mean		SD		F
	HP	LP	HP	LP	
HIGHLIGHT_SCORE	25.32	10.33	4.79	8.33	10.64**

***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$; HP=High Performance, LP=Low Performance

The average of highlight scores for both groups through 11 weeks are represented in Figure 3. The group with high learning performance consistently achieved a better highlight score than the group with low learning performance except for the last week since that week was for the final exam and no slides were uploaded. The first week was the introductory week in which the instructor introduced the basic concepts about the course. Both groups were able to achieve a good highlight score. It is observed that when the difficulty level of lectures increased as the week goes by, the highlight scores for both groups decreased. After week 4, both groups were more familiar with the courses and were able to highlight the key concepts. As both groups improved their highlight scores, the gap also enlarges. This indicates that the more concepts students learned, the more likely that students who gain a high highlight score can achieve a high learning performance. Therefore, we conclude that highlight quality strongly affects students' learning performance, and the impact is more significant as the learning period increases.

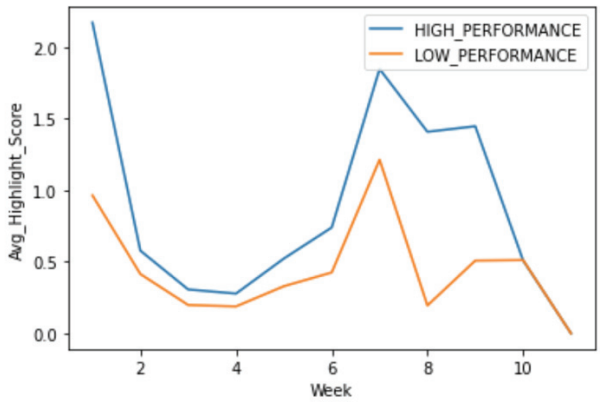

Figure 3. The average highlight score through 11 weeks.

4. Conclusion

In this research, we want to find a method to augment humans for automatically extracting the important concepts from learning materials and grading students' highlights since the accuracy of human grading may be affected when the number of highlights is too large and when graders have different knowledge levels. The summaries of four different text summarization techniques are compared with the gold standard answer from the instructor using BLEU 1, BLEU 2, BLEU 3, BLEU 4, METEOR, and ROUGE. The results show that when the content of slides consists of phrases or incomplete sentences, RAKE outperforms other techniques including deep learning algorithms. Then we use the summaries from RAKE as reference to calculate the highlight score using BLEU 1 considering most highlights contain only a few words. Finally, we explored whether highlight quality has influence on students' learning performance. The Spearman correlation between highlight quality and learning performance is 0.75 with P-value less than 0.001, which indicates that highlight quality is highly correlated to learning performance. The students were assigned into two groups. Since both groups show a normal distribution, a one-way ANOVA test is performed. The results show that the mean of highlight scores in high learning performance group is higher than the low learning performance groups at a statistical significance level. We further found that as students learned more concepts, the difference in highlight

score between two groups increased, which means students who achieve high learning performance are more likely to identify the key concepts from lectures.

In future work, we will apply the deep learning extraction model, BERT, on lectures that consist of full text, and expect to achieve a decent performance. Also, we will integrate this model into BoolRoll, provide feedback to students with the key concepts highlighted by the model, and investigate model's effectiveness by measuring students' learning performance.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innovation Program on AI and Big Data 18102059-0.

References

Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Flanagan, B., & Ogata, H. (2017, November). Integration of learning analytics research and production systems while protecting privacy. In The 25th International Conference on Computers in Education, Christchurch, New Zealand (pp. 333-338).

Lin, C. Y. (2004, June). Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough?. In NTCIR.

Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004. pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (July 2004)

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015, January). E-Book-based learning analytics in university education. In International conference on computer in education (ICCE 2015) (pp. 401-406).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. Text mining: applications and theory, 1, 1-20.