Generating Student Progress Reports based on Keywords

Shumpei KOBASHI* **& Tsunenori MINE**
Kyushu University, Fukuoka, Japan
*kobashi.shumpei@m.ait.kyushu-u.ac.jp

Abstract: In this paper, we propose a method that automatically generates a student learning status report based on keywords given by instructors at cram schools to reduce their burden on writing the report. For selecting sentences to generate the report, we propose two methods: Seq2Seq-based and Information Retrieval (IR)-based methods. The Seq2Seq-based method uses a Seq2Seq model to generate sentences using keywords given by the instructors. The IR-based method uses OkapiBM25 to select sentences from those written by the instructors based on the keywords. We conducted extensive experiments to evaluate the two methods on a test set of 197,493 sentences. The experimental results show that the Seq2Seq method generates more suitable sentences as the report than the IR-based method. Adding the attention mechanism to the Seq2Seq method further improved the performance of the Seq2Seq method. Considering the above experimental results, we discussed the generation of the lecturer report by keywords.

Keywords: Natural language generation, keywords, education support,

1. Introduction

In many cram schools, instructors are required to write reports on students after each class. This report can be used for encouraging students to study and sometimes for handover between instructors. Meanwhile, writing reports increases the instructors' burden to think about and write the reports each time, which may lead to the deterioration of sentences in the reports. Consequently, it may increase mass-produced reports with less meaningful contents that are not suitable for student learning. Therefore, it is an important issue for both students and instructors to automatically or semi-automatically generate meaningful reports.

In this study, we propose methods to automatically generate instructor reports based on keywords, where we assume that the instructors can select appropriate keywords considering the status of student learning. We propose two methods: a sequence to sequence (Seq2Seq)-based and an Information Retrieval (IR)-based method. The Seq2Seq-based method uses a model called the Seq2Seq model which consists of an encoder and a decoder (Sutskever et al., 2014); the encoder transforms a sequence, a set of keywords in this study, to a latent vector z mapped on a latent space and the decoder receives z and decodes it to another sequence, the report sentence in this study. The IR-based method uses OkapiBM25 (Robertson et al., 1995) and selects sentences from those written by the instructors based on the keywords. In this study, we conduct extensive experiments on a test set of 197,493 sentences given by the instructors at actual cram schools to examine the performance of the two methods. The experimental results show that the Seq2Seq-based method generates more proper sentences as the reports than the IR-based method. Adding the attention mechanism to the Seq2Seq-based method further improved the performance of the Seq2Seq method.

The main contributions in this study are as follows: (1) we pointed out the necessity to automatically generate report sentences so as to reduce instructors' burden, (2) we proposed two methods: Seq2Seq-based and IR-based methods to generate the sentences, and (3) we conducted extensive experiments on the real dataset of report sentences written by instructors at cram schools to compare the performance of the two methods.

In what follows, Section 2 shows literature reviews and discusses the position of this study; Section 3 explains proposed models; Section 4 discusses experimental results. Finally, we conclude and discuss our future work.

2. Related Work

Research and development on systems to support instructors in educational institutions have been conducted for a long time. The duties of instructors include teaching students, preparing questions, and grading tests. Gutl et al. (Gutl et al., 2011) survey and present various types of research on the automatic creation of test items. The latest automatic generation of test questions has been shown to be comparable in quality to human-created questions, contributing to the elimination of the need for instructors to create test questions. The purpose of this is to reduce the time required to create question materials, i.e., to reduce the burden on the instructor. Liang et al. (Liang et al., 2018) proposed a neural network-based Automated Essay Scoring (ASE) model to automatically grade essays so as to reduce the manual workload of instructors and to provide rapid feedback on learning. Considering the effectiveness of textual feedback and its human burden, Lu et al. (Lu et. al., 2021) have implemented an ASE model that combines word-embedding and a deep learning model, and then developed a text-based automatic feedback system using the Constrained Metropolis-Hastings Sampling sentence paraphrase unsupervised learning. Also, Malik et al. (Malik et al., 2021) introduced a generative grading system that provides nuanced and interpretable feedback by modeling the student's response process and learning the student's reasoning process. This system showed promising results across multiple modalities and domains. The research described above aims to reduce the burden on instructors by fully automating their work. In other words, we can say that we are developing robot (AI) instructors. However, Parab (Parab, 2020) shows that human instructors are more comfortable for students than robot instructors, and that it is beneficial for the system to support human instructors in their daily work. Based on the above, this research aims to support the daily work of instructors, namely "report generation" by using language processing technologies. Therefore, the goal in this study is not to completely automate the lecturers' work, but to reduce the workload of the lecturers while placing emphasis on the knowledge and insights of the human lecturers.

3. Proposed Methods to Generate Report Sentences based on Keywords

3.1 Proposed Methods

In this study, we propose two methods for generating sentences from keywords based on report sentences written by instructors at cram schools: one is the IR-based method, and the other is the Seq2Seq-based method. The IR-based method searches for past report sentences written by instructors using keywords, and selects the report with the highest rank. The Seq2Seq-based method learns a model to convert keywords to a report sentence based on report sentences written by instructors, and uses the model to generate a report sentence from keywords. Ideally, a comparison experiment between the two methods should be conducted using the following two sets of data: report sentences that instructors actually wrote and keywords that the instructors wanted to use to generate the report sentences. However, we can only use the data of the actual report sentences written by the instructors, which we call "original report data." Therefore, assuming the target report sentences would include keywords given by instructors, we extracted the keywords corresponding to each report sentence from the original report data and also assumed that the extracted keywords can be regarded as the keywords given by the instructors. Using the two sets of data: the original report data and keywords extracted from the original report data, we invent a task if report sentences in the original report data can be generated from the keywords extracted from the original report data, and evaluate the performance of the two proposed methods.

3.2 Keyword Extraction from Report Sentence

In this study, we use 197,493 report sentences provided by actual cram schools, which were written by instructors for each class regarding the learning status of their students. The average number of words per sentence is 15.6, which is shorter than a typical sentence. Keywords are extracted from each report sentence using Term Frequency and Inverse Documentation Frequency (tf-idf) weights, which are commonly used in keyword extraction. The specific procedure of keyword extraction is described as follows:

1. We apply morphological analysis to each sentence and divide it into morphemes. We use "MeCab"[1] as the morphological analyzer.
2. To calculate the tf-idf weights of the morphological data, we use TfidfVectorizer in the feature_extraction module of scikit-learn[2]. The tf-idf weights are calculated for the data excluding auxiliary verbs, particles, conjunctions, and interjection because these parts of speech are rarely used as keywords.
3. Words with tf-idf weight values above the average are adopted as keywords of the report sentences. The average number of keywords per sentence is 5.14. This means that about 1/3 of the words in the original sentence were extracted as keywords.

3.3 Overview of the Comparison Experiment

To evaluate the performance of the two models built by the two proposed methods on generating or selecting sentences, we conduct 10-fold cross-validation whose overview is shown in *Figure 1*. The training and test data are divided into 9 to 1 and the model is built from the training data, and the test data is used to evaluate the model built in the training phase, and the performance of the model is evaluated using the evaluation metrics described below. This cycle is repeated 10 times and we take average of the results.

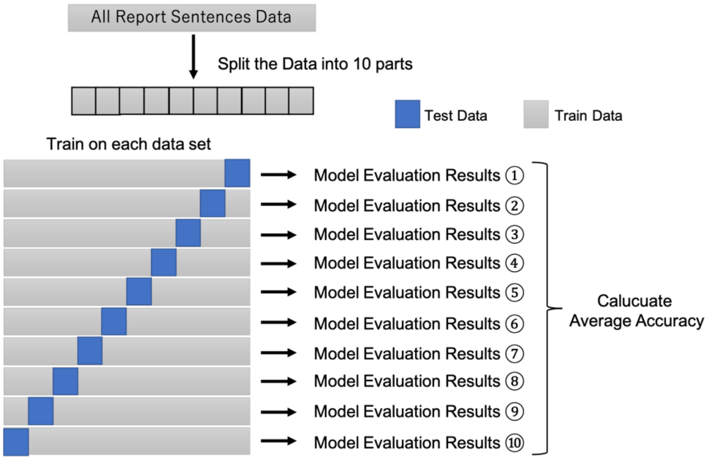
All Report Sentences Data
Split the Data into 10 parts
Test Data
Train Data
Train on each data set
Model Evaluation Results ①
Model Evaluation Results ②
Model Evaluation Results ③
Model Evaluation Results ④
Model Evaluation Results ⑤
Model Evaluation Results ⑥
Model Evaluation Results ⑦
Model Evaluation Results ⑧
Model Evaluation Results ⑨
Model Evaluation Results ⑩
Calucuate
Average Accuracy

Figure 1. Overview of the 10-Fold Cross-Validation.

As metrics to evaluate sentences generated by each model, we use BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), which is often used in machine translation tasks, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), which is used in summarization tasks, and CR (Content-Rrate), which indicates how many keywords given as input data are included in the generated sentences. If CR is 0.5, it means that the sentence contains the half of the keywords given as input.

3.4 Information Retrieval-Based Method

Our IR-based method uses OkapiBM25 (BM25 for short), which is a well-known and commonly used method in the field of information retrieval and is expected to be more accurate than simply using tf-idf values. In this study, we rank the report sentences by the BM25 score based on the input keywords and return the highest ranked report sentences. In other words, this method selects sentences rather than generates them. *Figure 2* illustrates the process of generating sentences based on keywords using BM25. In the IR-based method, the sentence with the highest BM25 score is selected from the training data as output. In a simple way, we need to calculate the BM25 scores of all sentences in the training

[1] http://taku910.github.io/mecab/
[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

data. However, the BM25 score is inevitably higher for sentences that contain many words that are input keywords. Therefore, in this study, we first select sentences containing the most input keywords as candidate sentences, calculate the scores of the candidate sentences, and select the sentence with the highest score as the generated sentence.

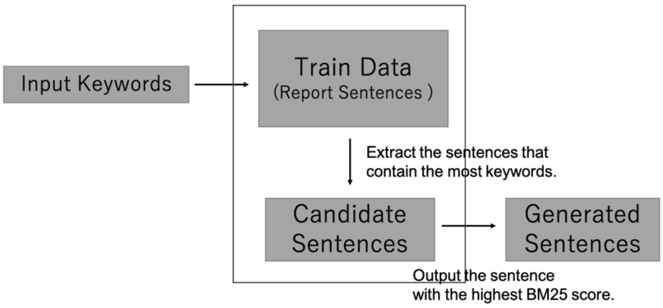
Input Keywords
Train Data
(Report Sentences)
Extract the sentences that contain the most keywords.
Candidate Sentences
Generated Sentences
Output the sentence with the highest BM25 score.

Figure 2. The Process of Sentence Generation by IR-based Method.

3.5 Seq2Seq-based Method

Seq2Seq is a model that uses a neural network to convert a sequence data to another sequence data. Seq2Seq is composed of two mechanisms: Encoder and Decoder. LSTM (Long short-term memory) (Gers et al., 1999) is often used for each mechanism. Encoder takes a sequence as input and maps it to a fixed-dimensional vector. The decoder decodes the fixed-dimensional vector output by the encoder to the target sequence. In this experiment, the model is trained to encode the input keywords and decode them into the original report sentences. Here, there are two methods for sentence generation: probabilistic generation (PG) and deterministic generation (DG); PG chooses words according to a probability distribution, and DG uniquely chooses the most probable word. We respectively call the two methods, the Seq2Seq with PG and Seq2Seq with DG methods.

We compare the performance of these methods on generating report sentences. *Figures 3* showss the overviews of the Seq2Seq model during training and evaluation phases, respectively. During training, [KEYWORDS] feeds Instructor's Report Sentence-*i* in the training data and extracts Keywords-*i* that are paired with the Instructor's Report Sentence-*i* using the method described in Section 3.1. Giving the Instructor's Report Sentence-*i* and Keywords-*i* to [SEQ2SEQ] as input, [SEQ2SEQ] adjusts the parameters. By iterating this process, [Trained SEQ2SEQ] is built. When evaluating the model, the Instructor's Report Sentence-*j* is taken from the test data and given to [KEYWORDS]. Keywords-*j* are extracted in the same way as during training. With this Keywords-*j* as input, [Trained SEQ2SEQ] generates the Generated Instructor's Report Sentence-*j*. To evaluate the model, the Instructor's Report Sentence-*j*, Keywords-*j*, and the Generated Instructor's Report Sentence-*j* are given to [JUDGE], and the evaluation score (Score) is calculated according to the metrics: BLEU, ROUGE and CR.

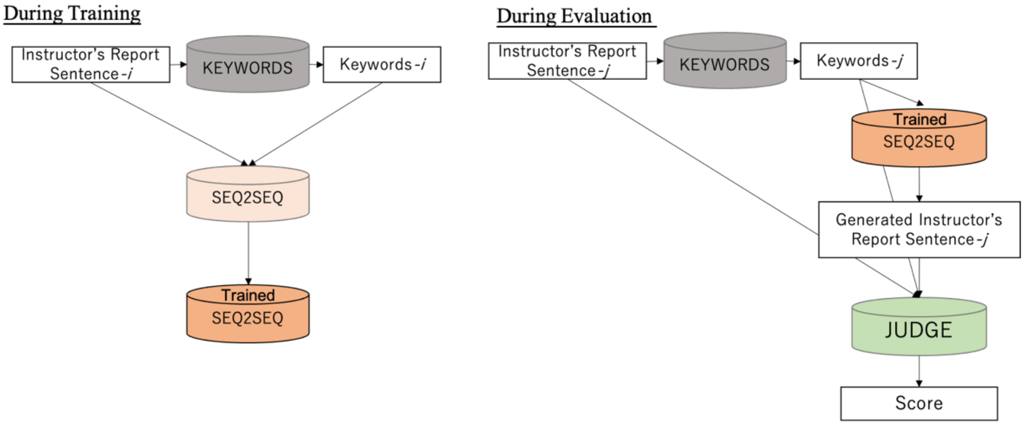

Figure 3. The Seq2Seq Model.

4. Experiments

4.1 Comparison between IR-based and Seq2Seq-based Methods

The experimental results are shown in *Figure 4*. For the content-rate, the IR-based method had the highest accuracy, and for the other metrics, the Seq2Seq with DG method had the highest scores. The sentences generated (selected) by the IR-based method tend to include keywords appearing separately, but not connectively, which reduces the rate of n-grams, makes BLEU and ROUGE scores lower, and often generates sentences with unintended contexts. As the BM25 score is higher, the greater number of keywords are included, which selects longer sentences more likely. At the same time, this makes it difficult to generate short and concise sentences. As a result, the Seq2Seq-based method performs better than the IR-based method when generating concise and/or abstract sentences, which are the most part of the test data. On the other hand, the Seq2Seq-based method is not good at generating concrete and complex long sentences, which are well-suited for the IR-based method.

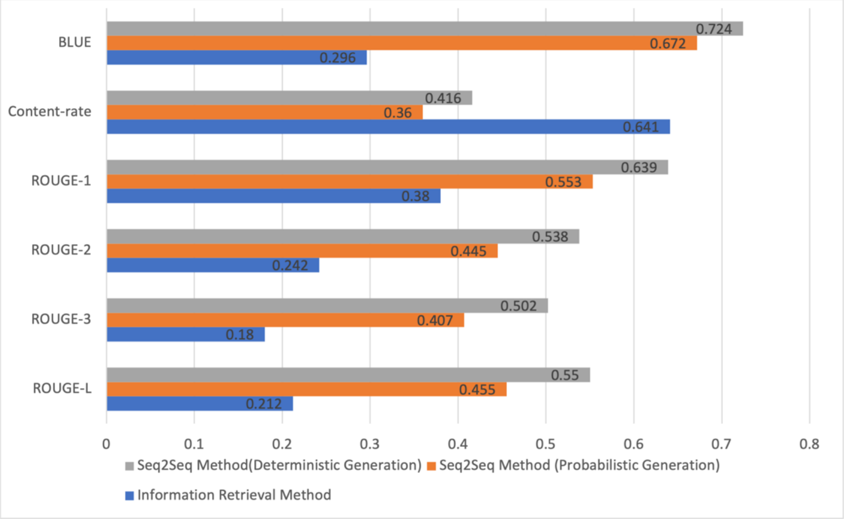

Figure 4. The Comparison Results of the Two Methods.

4.2 Improving Performance of Seq2Seq-based Method using Attention Mechanism

As mentioned in the previous section, the Seq2Seq with DG method had the highest accuracy. In this section, we evaluate the effect of Attention mechanism (Vaswani et al., 2017), which is a mechanism that can directly refer to the information of the input sequence at the time of decoding, and can consider the length of the input sequence at the encoder side. Here, this Attention refers to soft Attention. We add the attention mechanism to the Seq2Seq with DG and call the Seq2Seq with Attention method. In contrast, we call the Seq2Seq with DG, but without Attention mechanism, the Seq2Seq without Attention method. We compare the Seq2Seq with and without Attention methods and show the results in *Table 1*. As we can see, the Seq2Seq with Attention method has better performance on ROUGEs, but slightly worse on BLEU. This shows that the Attention mechanisms work effectively in generating report sentences from keywords. Also, *Table 2* shows the specific results of the sentences generated by the Seq2Seq with Attention.

Table 1. *Comparison of Seq2Seq with and without Attention Methods*

	Seq2Seq without Attention	Seq2Seq with Attention
BLEU	***0.724***	0.716
ROUGE-1	0639	***0.725***
ROUGE-2	0.538	***0.620***
ROUGE-3	0.502	***0.565***
ROUGE-L	0.550	***0.608***

Table 2. *Report sentences Generated by the Seq2Seq with Attention*

Keywords	Generated sentence
単語　ミス　有り (word errors exist)	単語のミスが少し有りますが、少しずつ理解してくれています （There are a few word errors, but you’re slowly getting the hang of it.）
授業　集中　できる （class concentrate can）	授業中、集中して問題に取り組んでくれました （You concentrated on the problem during the class）

5. Conclusion

This paper discussed report generation models for instructors who have to write reports on learning status of their students. Using the models, the instructors can obtain student progress report by simply inputting keywords instead of writing many sentences. This can greatly reduce instructors’ burden on generating report sentences. In this study we proposed two sentence generation methods: the IR-based method using OkapiBM25 and the Seq2Seq-based method, and compared their performance by conducting experiments on a dataset consisting of about 200,000 report sentences written by the instructors at cram schools. Experimental results show the Seq2Seq-based method outperforms the IR-based method and adding the attention mechanism to the Seq2Seq-based method had effects on improving the performance of generating sentences. However, our keyword-based sentence generation method using the Seq2Seq-based method still has some weak points, especially on generating concrete and detail sentences, which will greatly improve the usability of this model. We will commit to tackle this problem and report it elsewhere in near future. In addition, the system that actually uses Seq2Seq with Attention is currently being highly evaluated and used by people involved in the cram schools.

Acknowledgements

This work was supported in part by e-sia Corporation and by Grant-in-Aid for Scientific Research proposal numbers (JP21H00907, JP20H01728, JP20H04300, JP19KK0257, JP18K18656). We would like to express our deepest gratitude to them.

Reference

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. Nist Special Publication Sp, 109, 109.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215.

Gutl, C., Lankmayr, K., Weinhofer, J., & Hofler, M. (2011). Enhanced Automatic Question Creator--EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education. *Electronic Journal of e-Learning*, *9*(1), 23-38.

Liang, G., On, B. W., Jeong, D., Kim, H. C., & Choi, G. S. (2018). Automated essay scoring: A siamese bidirectional LSTM neural network architecture. *Symmetry*, *10*(12), 682.

Lu, C., & Cutumisu, M. (2021). Integrating Deep Learning into An Automated Feedback Generation System for Automated Essay Scoring. EDM 2021

Malik, A., Wu, M., Vasavada, V., Song, J., Coots, M., Mitchell, J., ... & Piech, C. (2021). Generative Grading: Near Human-level Accuracy for Automated Feedback on Richly Structured Problems. EDM 2021

Parab, A. K. (2020). Artificial Intelligence in Education: Teacher and Teacher Assistant Improve Learning Process. International Journal for Research in Applied Science & Engineering Technology, 8(11), 608-612.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. ACL 2002, 311-318.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Text summarization branches out, 74-81.

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.