Feedback comment generation using predicted grammatical terms

Kunitaka Jimichi[1], **Kotaro Funakoshi**[1], **Manabu Okumura**[1,2]
[1]Tokyo Institute of Technology [2]RIKEN Center for Advanced Intelligence Project
{kunitaka,funakoshi,oku}@lr.pi.titech.ac.jp

Abstract

The purpose of feedback comment generation is to provide useful feedback comments for a wide range of errors in learners' essays from a language learning perspective. Since it is difficult to obtain appropriate comments at a practical level with rule-based or retrieval-based methods, we explore neural-based generative methods with pre-trained models. We further assume the effectiveness of considering grammatical terms in generating feedback comments. Specifically, this paper proposes T5-based models using predicted grammatical terms, submitted to FCG GenChal, and presents their results. By using correct grammatical terms, our model could improve the BLEU score by 19.0 points, compared with the baseline T5 without grammatical terms on the development dataset. Furthermore, by using predicted grammatical terms, our model could improve the manual evaluation score by 2.33 points, compared with the baseline T5 without grammatical terms on the test dataset.

1 Introduction

Feedback comment generation (FCG) for writing studies is a task to generate explanations of why ungrammatical sentences written by language learners are incorrect and on what basis a correction was made. Related research has been mainly conducted on the basis of a dataset published by Nagata (2019). However, the accuracy, which is based on *manual evaluation* (ME), so far is insufficient for real-world use. One reason is that the data in the dataset are not necessarily sufficient to cover all error types and feedback comments. Since annotating feedback comments would require specialized knowledge in writing, constructing larger datasets is time-consuming and labour-intensive.

FCG GenChal (Nagata et al., 2021) targets the generation of feedback comments for prepositional errors. An example feedback comment for the prepositional error is shown in the following.

> Target sentence: I agree <u>on</u> the idea.
>
> Feedback comment: «Agree on» indicates that more than one person agrees on a certain matter. <verb> «agree» to find the <preposition> used to indicate that the same idea is shared.

Here, the words marked by <> are *grammatical terms* (GTs). Underlined words are the target word with an error that the feedback comment is generated for.

To achieve this task, methods using an Encoder-Decoder model, which generates feedback comments from scratch, are considered effective because they can deal with various learners' error types even in unsophisticated settings. Therefore, it is a good idea to develop the Encoder-Decoder model for generating feedback comments.

State-of-the-art (SOTA) results have been achieved in various natural language processing tasks by using pre-trained models. However, there has been no previous work on FCG using pre-trained models. A logical concern is what level of accuracy the SOTA pre-trained models can achieve in FCG. In grammatical error correction, which is highly related to FCG, methods using T5 (Raffel et al., 2020) achieve good results. Thus, in this work, we also utilize T5 for FCG.

The system can more easily generate a feedback comment when some words in the target feedback comment sentence to be generated are known. One of such clue words for better generating feedback comments might be grammatical terms (GTs), since commonly used GTs are limited, and it might be possible to predict and use them to generate feedback comments. However, no research has focused on GTs for FCG.

Therefore, we propose the following procedure for generating feedback comments in this study. First, GTs to be used in the feedback comment are selected. As several GTs are used in a feedback

comment, this becomes a multi-label classification task. Next, the selected GTs are used to generate the feedback comment.

To predict GTs, we use RoBERTa (Liu et al., 2019) because RoBERTa often achieves better accuracy than other pre-trained models in the multi-label classification task. T5 is then used to generate feedback comments since it can be used for text-to-text tasks.

The contributions of this research are therefore as follows:

- We investigate the extent to which the use of GTs improves the ME in FCG by using T5.
- We demonstrate the use of correct GTs using the T5$_{base}$ model improves the BLEU (Papineni et al., 2002) score by 19.0 points on the development dataset, and predicted GTs using the T5$_{base}$ model improves the ME score by 2.33 points on the test dataset.

2 Related work

Grammatical error correction is closely related to the FCG task. Rothe et al. (2021) have achieved a high accuracy in grammatical error correction by using a pre-trained generative language model, T5. This suggests that FCG could also be handled by T5.

A survey (Hanawa et al., 2021) of the methods used in the FCG task investigated three methods: retrieval-based, retrieve-and-edit, and simple generation. The survey shows that the simple generation method performs best in generating feedback comments for prepositional errors and the retrieval-based method alone cannot cope with various errors present in the training examples in generating feedback comments.

In generating feedback comments using a generative model, prompting the model with the predicted GTs corresponding to the target error is likely to guide the direction for the generation. However, to the best of our knowledge, there have been no studies taking such an approach to FCG.

3 Grammatical term prediction

3.1 Task definition and notations

This section describes the prediction task of GTs. To define the task formally, we introduce the following symbols. The learners' sentence, its length (the number of tokens), and the i-th token are denoted by S, N and w_i, respectively. That is, $S = w_1, ..., w_i, ..., w_N$. The span where to comment is denoted by s,[1] which indicates the position of several consecutive words. The task input is defined as $x = (S, s)$. The GTs and their number in the reference feedback comment y corresponding to S are denoted by T and M, respectively. That is, $T = t_1, ..., t_i, ..., t_M$. Here, T is sorted in lexicographic order. $M = 0$ means that y does not contain any GTs. The grammatical term prediction task is to predict T given x.

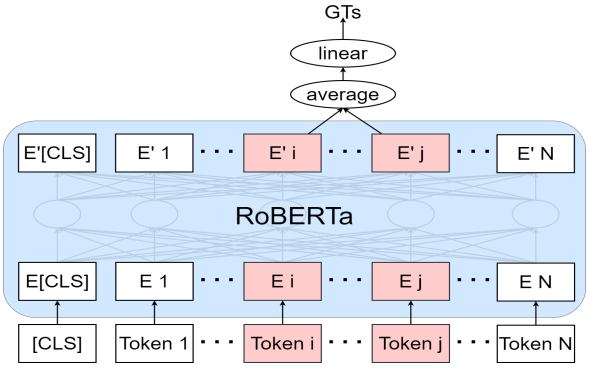

Figure 1: Schematic diagram of multi-label classification for GTs with RoBERTa.

3.2 Prediction method

We use RoBERTa as the base model. Each input word in S is transformed into the corresponding embedding, which is then transformed into a context-aware embedding by RoBERTa. The embeddings of the words in s are then averaged and fed into a linear layer. The GTs whose probabilities are above a certain threshold θ are to be included as the prompt used in feedback comment generation, described in the next section. An overview of the model is shown in Figure 1.

4 Feedback comment generation

In FCG, T5 is used as the base model to predict $\hat{y}$ given x and T. The input sequence to T5 is given in the following format:

fbc: w_1 ... [... w_i ...] ... w_N <GTs> : t_1 ... t_i ... t_M

Here, a special token "fbc:", which stands for a *feedback comment*, is used as a prompt to train the T5 model. The target span s is marked by square brackets. Predicted GTs for S are listed after S with another special token "<GTs>".

[1]These spans are referred to as *offsets* in the shared task.

Data	Instances	Sent toks.	Com toks.
Train	4868	110906	127439
Dev.	170	3142	4516
Test	215	4446	-

Table 1: Statistics of the dataset. Instances, sent toks., and com toks. denote the number of instances, tokens in learners' sentences and tokens in reference feedback comments, respectively. The information for the reference feedback comments in the test dataset is not included because FCG GenChal participants cannot get access to them.

Data	Method	BLEU	ME
Dev.	$T5_{small}$	47.6	-
	$T5_{small}$ + predicted GTs (top-10)	45.9	-
	$T5_{base}$	49.6	-
	$T5_{base}$ + predicted GTs (top-10)	49.0	-
	$T5_{small}$ + correct GTs (top-10)	61.0	-
	$T5_{small}$ + correct GTs (full)	64.7	-
	$T5_{base}$ + correct GTs (top-10)	63.0	-
	$T5_{base}$ + correct GTs (full)	68.6	-
Test	Baseline system	33.4	31.16
	$T5_{small}$ + Predicted GTs (top-10)	46.0	56.28
	$T5_{base}$	-	58.14
	$T5_{base}$ + Predicted GTs (top-10)	-	60.47

Table 2: Feedback comment generation results on the development and test datasets. ME: manual evaluation.

5 Experiments

5.1 Dataset and metrics

We used the shared task data (Nagata et al., 2021). The data was originally divided into three sets, which are training, development, and test sets, by the FCG Organizers. The statistics of the dataset are shown in Table 1. The training dataset was used for fine-tuning RoBERTa and T5. The development and test datasets were used for evaluating the methods for FCG.

For grammatical term prediction, considering the relatively small size of the data, we used only the top 10 most frequent GTs and discarded the others. The top ten most frequent GTs are <preposition>, <verb>, <noun>, <object>, <transitive verb>, <intransitive verb>, <to-infinitive>, <noun phrase>, <adjective>, and <to infinitive>.[2] For evaluation of grammatical term prediction, only the development dataset was used.

The precision, recall, and F1 of ME, and BLEU (N=4) are the evaluation measures used in accordance with FCG GenChal. The ME scores are based on the human judgment of whether a system output is correct or not. More details are found on the page of the FCG GenChal task description.[3] Since all values of precision, recall, and F1 are the same, only F1 is listed in the table. GTs (full) means the case using all GTs rather than only top-10 GTs.

To evaluate grammatical term prediction, we used exact match ratio (EMR), which indicates the percentage of instances that have all their labels classified correctly. In addition, we used micro averaged precision, recall, and F1 with GTs as a unit.

5.2 Hyperparameters

Grammatical term prediction The RoBERTa model used in the experiments was roberta-large.[4] We tuned the learning rate from 0.00001, 0.00003, and 0.0001, and the threshold θ with the highest EMR on the training dataset. The learning rate was fixed to 0.00003 and the threshold θ was fixed to 0.68604184.

AdamW was used as the optimisation function. A batch size of 8 was used and a drop-out rate of 0.1 was used for each linear layer. The maximum sentence length was set to 256. We added one linear layer not included in RoBERTa, with a size of 1024×10. The hidden layer size of roberta-large is 1024 and the number of GT types is 10. No drop-out was applied to the linear layer. Each word was lowercased. The number of epochs used for training was 5. We applied a weight to each GT label when calculating the loss for it. We used the inverse document frequency (IDF) of each label as the weight and calculated it within the training dataset.

Feedback comment generation The T5 model used in the experiments was $T5_{small}$ and $T5_{base}$.[5] A learning rate of 0.0001 was used. AdamW was used as the optimisation function. A batch size of 8 was used and a drop-out rate of 0.1 was used for each linear layer. The maximum sentence length was set to 512. Each word was lowercased. The number of epochs used for training was 50. "GTs", "``", "``", "´´", "«", "»", "<" and ">" were added to the T5 dictionary as special tokens.

[2] In these experiments, <to-infinitive> and <to infinitive> were used as separate terms.

[3] https://fcg.sharedtask.org/task/

[4] https://huggingface.co/roberta-large

[5] https://huggingface.co/docs/transformers/model_doc/t5

EMR	P	R	F
8.23	44.68	42.61	42.65

Table 3: Grammatical term prediction results on the development dataset. EMR=exact match ratio, P=micro averaged precision, R=micro averaged recall, and F=micro averaged F1-measure.

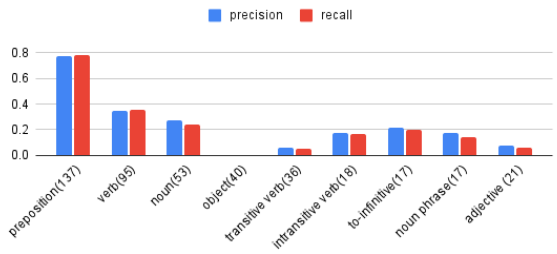

Figure 2: Precision and recall for each label in grammatical term prediction; blue bars are precision and red bars are recall. The number in brackets next to the label indicates the number of occurrences on the development dataset.

Reference	The <compound preposition> «because of» should be followed by a <noun>. «Of» is unnecessary when a <clause> follows.
Our Model	The <compound preposition> «because of» should be followed by a <noun>. «Of» is unnecessary when a <clause> follows.
Predicted GTs	<noun> <preposition>
Reference	When a <noun> is qualified by another <noun> that follows, a <preposition> is necessary between the two nouns. Think of the most common <preposition> of association.
Our Model	The <preposition> to indicate the direction of negative influence is missing. Look up the use of the <noun> «future» in the dictionary and add the appropriate <preposition>
Predicted GTs	(*no output*)

Table 4: Case study: Two qualitative examples; one for which the model predicted the GTs <noun> and <preposition>, and one for which the model failed to predict any GTs (no output). In the top example, only the GT <noun> was successfully predicted and the generated feedback comment was correct. In the bottom example, no GTs were predicted and the generated feedback comment was incorrect.

5.3 Results

The results for the FCG are shown in Table 2. The table shows that T5$_{small}$ using correct GTs (top-10 and full) improves the BLEU score by 13.4 and 17.1 points on the development dataset, respectively. We also found that T5$_{base}$ using correct GTs (top-10 and full) improves the BLEU score by 13.4 and 19.0 points on the development dataset, respectively. Furthermore, T5$_{base}$ using predicted GTs improves the ME score by 2.33 points on the test dataset. This indicates that incorporating predicted GTs in FCG is beneficial for T5.

The results of grammatical term prediction are shown in Table 3. The table shows the multi-labelling performance of the submitted model on the development dataset.

We independently investigated the precision and recall for each GT in the top 10 most frequent words. The results are shown in Figure 2 (Since <to infinitive> was not present in the development dataset, it was not included in the target GTs to be predicted and is excluded from the figure). The figure shows that the precision and recall for each GT do not depend on its frequency in the training dataset. The highest precision and recall are for <preposition>, followed by <verb>, <noun>, and <to-infinitive>. The high prediction performance for <to-infinitive> would be related to the ease of predicting the error type in an English sentence. When we find a case of two consecutive verbs or 'to infinitive' + the 'ing' form of a verb in a sentence, we can simply determine there is an error in it.

6 Case study

We investigated whether our model could generate correct feedback comments with the predicted GTs in the development dataset. Table 4 shows examples where our model produced correct and incorrect feedback comments in the top and bottom rows, respectively. In the top example, while our model correctly predicted <noun>, it also incorrectly predicted <preposition>, a GT similar to the correct <compound preposition>. In the bottom example, it did not predict any GTs, and as a result, gains no benefit from them and generates an incorrect feedback comment.

7 Conclusion

We explored neural methods for FCG using pre-trained models. In this study, we showed predicting the GTs and using them in generating feedback comments can be useful for feedback comment generation with T5. The results also suggested that further improvement in grammatical term prediction would be beneficial for FCG.

Acknowledgements

The authors are grateful to Prof. Ryo Nagata in Konan University for suggesting the topic addressed in this paper. We also thank the FCG Organizers for sharing their dataset and hosting the FCG GenChal for us.

References

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring methods for generating feedback comments for writing learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9719–9730, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.