Shared Task on Feedback Comment Generation for Language Learners

Ryo Nagata
Konan University, Japan
JST, PRESTO, Japan
nagata-genchal@ml.hyogo-u.ac.jp.

Masato Hagiwara
Octanove Labs, USA
masato@octanove.com

Kazuaki Hanawa
RIKEN, Japan
Tohoku University, Japan
kazuaki.hanawa@riken.jp

Masato Mita
RIKEN, Japan
Tohoku University, Japan
masato.mita@riken.jp

Artem Chernodub
Grammarly
artem.chernodub@grammarly.com

Olena Nahorna
Grammarly
olena.nahorna@grammarly.com

Abstract

In this paper, we propose a generation challenge called *Feedback comment generation for language learners*. It is a task where given a text and a span, a system generates, for the span, an explanatory note that helps the writer (language learner) improve their writing skills. The motivations for this challenge are: (i) practically, it will be beneficial for both language learners and teachers if a computer-assisted language learning system can provide feedback comments just as human teachers do; (ii) theoretically, feedback comment generation for language learners has a mixed aspect of other generation tasks together with its unique features and it will be interesting to explore what kind of generation method is effective against what kind of writing rule. To this end, we have created a dataset and developed baseline systems to estimate baseline performance. With these preparations, we propose a generation challenge of feedback comment generation.

1 Introduction

Feedback comment generation for language learners is a task where given a text and a span, a system generates, for the span, an explanatory note that helps the writer (language learner) improve their writing skills (for convenience of explanation, the task will be abbreviated as *feedback comment generation*, hereafter). The target language(s) can be any language, but we limit ourselves to English input texts and English feedback comments in this challenge. As shown in Figure 1, a feedback comment is typically made about erroneous, unnatural, or problematic words in a given text so that the writer can understand why the present form is not good together with the underlying rule.

In this regard, feedback comment generation is related to grammatical error detection and correction. In many cases, however, it is not enough to just point out an error with its correct form in order to help language learners with writing learning. Instead, it is often essential for them to explain underlying rules, which makes the task different from the conventional grammatical error detection/correction. In other words, it is essential in feedback comment generation to include more information than grammatical error detection/correction provide.

At the same time, unconstrained generation would make the task infeasible in terms of system development and evaluation. With this in mind, we set some constrains to the task to be explored in this generation challenge as described in Section 2. For example, the input is limited to a sentence (and a span) instead of a text.

The motivations for this challenge are as follows. A practical motivation is already mentioned above. It will be useful if a computer-assisted language learning system can provide feedback comments just as human teachers do. Theoretically, feedback comment generation has a mixed aspect of other generation tasks together with its unique features as described in Section 3. It will be interesting to explore what kind of technique is effective against what kind of writing rule.

One of the goals of this challenge is to reveal how well we can generate feedback comments with existing techniques. There is a wide variety of choices as generation methods that are applicable

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 320–324,
Aberdeen, Scotland, UK, 20-24 September 2021.

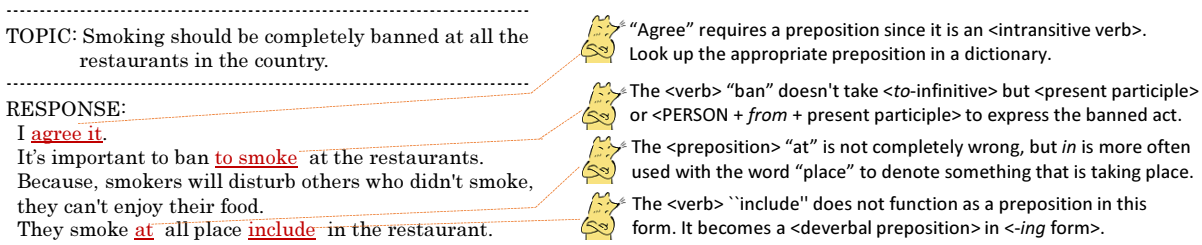

Figure 1: Example of Feedback Comments.

to this task. Nevertheless, they have not yet been explored (at least, much less than in other generation tasks). The generation challenge will enable the NLG community to evaluate and compare a range of techniques using the same dataset. Besides it will provide us with opportunities to develop new techniques.

Another goal is to shed a light on the difficulties in this task. This is going to be the first generation challenge of feedback comment generation as far as we are aware of. No one fully understands what is possible and impossible in the task. Holding this generation challenge will bring more insights into the task, which will in turn give new knowledge and experience to the NLG community.

Having said that, to make the task feasible within GenChal2021, we have prepared a dataset, evaluation metrics, and other necessities such as tools for this challenge as shown in Section 4. We have even developed baseline systems to estimate baseline performance. With these preparations, we propose a generation challenge of feedback comment generation.

2 Task Definition

2.1 General Definition

A unit of the input in feedback comment generation in general consists of a text and spans of the text. Spans, which are counted by 1-based index based on characters, correspond to where to comment. An example input text would be:

(1) *I agree it.*

as shown on the left-hand side of Figure 1. A span would be 3 to 10, which will be abbreviated as 3:10, hereafter.

The output for a span is a string that explains why the span is not good together with the underlying rule. To make the task different from grammatical error detection/correction, the output string has to contain more information than grammatical error detection/correction provide. In other words, just indicating the error position, the erroneous word(s), and/or the correct form are not enough as a valid feedback comment, of which details are discussed in Subsection 2.2.

2.2 Task Definition to Be Used

The above task definition is too general and abstract to be a practical one. For this reason, we put some constraints on it.

First, we limit the target only to preposition use as in the examples in Figure 1. It should be emphasized that the target includes missing prepositions, *to*-infinitives, and deverbal prepositions (e.g., *including*) in preposition use.

Second, we also limit the input to a narrower unit. Specifically, the input text always consists of only one sentence with one span. Also, they are pre-tokenized where tokens are separated by whitespace. For example, the first sentence in Figure 1 would give an input:

(2) *I agree it .* \t 3:10

where \t stands for the tab character. If a sentence contains more than one preposition error, it appears two or more times with different spans.

Under these settings, participants develop a system that automatically generates an appropriate feedback comment in English for an input sentence and a span. The length of a generated feedback comment should be less than 100 tokens. If a system cannot generate an appropriate feedback comment for a given span, it may generate the special token <NO_COMMENT>, which is not counted as a system output. This allows us to calculate recall, precision, and F_1 as explained below. An example output would be:

(3) *I agree it .* \t 3:10 \t **"agree" is an intransitive verb and thus it requires a preposition before its object.**

Also note that the input sentence and its span are included in the system output for evaluation convenience.

Evaluation is probably the hardest challenge in this task. We adopt automated and manual evaluation methods. In the former, we simply take BLEU between a system output and its corresponding reference (manually created feedback comment). In the latter, human evaluators examine whether a system output and its corresponding reference are equivalent in meaning. To be precise, a system output is regarded as appropriate if (1) it contains information similar to the reference and (2) it does not contain information that is irrelevant to the span; it may contain information that the reference does not contain as long as it is relevant to the span. This way of manual evaluation inevitably brings subjectivity to some extent. In practice, however, the results of a pilot study show that inter-evaluator agreement is considerably high as shown in Section 4.

The final manual evaluation measures are recall, precision, and F_1. Recall is defined as the number of appropriate system outputs divided by the number of target spans. Similarly, precision is defined as the number of appropriate system outputs divided by the number of system outputs where the special output <NO_COMMENT> is excluded. F_1 is the harmonic mean of recall and precision.

3 Related Work

Generally speaking, feedback comment generation is a task of text-to-text generation. It then can be abstractly regarded as a Machine Translation (MT) problem where the input text, which is written by a language learner, is translated into another text explaining writing rules. This implies that generation methods employed in MT or other related research areas may be effective in the present task. For example, feedback comments often refer to words and phrases appearing in the input text, and techniques for referring to words in the source text (e.g., copy mechanisms) will likely be beneficial.

Unlike MT, the equivalence between the source and target texts in meaning do not hold. Instead, the target text is a feedback comment that explains why the source is not good together with the underlying rule. From this point of view, the present task is related to research in dialogue systems.

Feedback comment generation has its unique aspects as well. It should be emphasized that a feedback comment is generated against a span (of the input text or sentence) whereas only a text (e.g., a sentence or utterance) is dealt with in other major text-to-text generation tasks such as MT and dialog systems. In consequence, feedback comment generation systems have to output different texts for the exact same source sentence, depending on given spans.

The source and target languages are also unique. In this challenge, both are English, but there is room for discussion whether they fall into the same language class. The former is learner English, and inevitably it contains erroneous/unnatural words. Even within correct sentences, grammar, expressions, and style are expected to be used differently from canonical English. This brings out further research questions related to the source and target languages. For example, which is the best setting of vocabularies — only one common vocabulary for the source and target, or one for each? Does a pre-trained general (or native) language model work well to model learner English? There are a number of unaddressed research questions like these.

Feedback comment generation is also related to grammatical error detection/correction. The state-of-the-art methods typically solve the problems as sequence labeling (e.g., Kaneko et al. (2017)) or MT with DNNs (e.g., Junczys-Dowmunt et al. (2018); Napoles and Callison-Burch (2017); Rothe et al. (2021)). Recently, a DNN-based sequence labeling method is combined with symbolic transformations (Omelianchuk et al., 2020), which can be a good source of information to generate feedback comments.

Some researchers (Kakegawa et al., 2000; McCoy et al., 1996; Nagata et al., 2014) made an attempt to develop rule-based methods for diagnosing errors in line with grammatical error correction. Researchers started to apply more modern techniques. Nagata (2019) showed that a neural-retrieval-based method was effective in preposition feedback comment generation. Lai and Chang (2019) proposed a method that used grammatical error correction and templates to generate detailed comments. Gkatzia et al. (2013) and Gkatzia et al. (2014) proposed methods for automatically choosing feedback templates based on learning history.

The availability of datasets for research in feedback comment generation has been increasing. Nagata (2019) released a dataset consisting of feed-

back comments on preposition use. They marked up erroneous prepositions and annotated them with feedback comments. Nagata et al. (2020) extended it to other grammatical errors and also other writing items such as discourse and lexical choice. Pilan et al. (2020) released a unique dataset where feedback comments on linking words were annotated.

4 Preparation

Based on the work (Nagata, 2019; Nagata et al., 2020), we created a new dataset for this generation challenge. The original texts are excerpts from the essays (written by learners of English) in ICNALE (Ishikawa, 2011). We had native speakers of English, who had experience in English teaching, manually annotated all preposition errors with feedback comments in English. Table 1 shows its statistics.

The dataset will be split into training, development, and test sets. Note that training and development sets consist of the whole essays, meaning that they contain all sentences no matter whether they contain feedback comments or not (i.e., error free essays are included in the sets). Also note that a sentence can be annotated with more than one feedback comment. In contrast, the test set only contains sentences with exactly one feedback comment each as described in Subsection 2.2. If a sentence contains more than one preposition error, it appears two or more times with different spans (in different lines). They will be provided for the participants in due course.

We also developed a Web-based submission system that takes system outputs the participants submit. The system checks the output format of the submission and calculate its score (i.e., BLEU).

We also implemented two baseline systems for this challenge. One is a deep neural network (DNN)-based retrieval system that retrieves the most similar instance to the input sentence and outputs it as a generation result. The other is a text generation system based on a DNN encoder-decoder with a copy mechanism.

As a pilot study, we tested them on the dataset (Nagata, 2019). For manual evaluation, we trained a professional annotator who had more than ten years of experience in English syntactic annotation and two years of experience in professional English writing teaching. The person and the first and third authors independently evaluated the generation results. They labeled each generated feedback comment as either *appropriate* or not, following the manner described in Subsection 2.2.

Corpus	ICNALE
Number of essays	1,884
Number of sentences	27,995
Number of tokens	473,815
Number of comments	5,237

Table 1: Statistics on Dataset.

As a result, it turned out that the retrieval system and the text generation system achieved an F_1 of 0.35 and 0.43, respectively[1]. Inter-evaluator agreements (Cohen's kappa coefficient) were considerably high, ranging from 0.86 to 0.92. These results imply that the present task is not easy one, but also not completely insolvable.

5 Organizers

The organizing group comprises six people as shown in the authors of this paper. All members have engaged in natural language research related to language learning and education for many years and some of them have organized several workshops and shared tasks in the domain.

The first author developed the dataset. The second author developed the submission system together with a Web page for this challenge. The two mainly designed this generation challenge with help from the other members. The third author implemented the baseline systems with the first author. They were also involved in the pilot manual evaluation.

All will be involved in the actual generation challenge including evaluation and paper publication. Although the trained professional evaluator is not included in the organizing group, the person will play a major role in manual evaluation.

6 Conclusions

In this paper, we have described a new generation challenge called *Feedback comment generation for language learners*. We have explored the task, describing the task definition, the related work, and the dataset to be used.

[1]The baseline systems are not designed to generate the special token <NO_COMMENT>, and they always output a feedback comment for a given span. Accordingly, it always holds that recall = precision =F_1.

References

Dimitra Gkatzia, Helen Hastie, Srinivasan Janarthanam, and Oliver Lemon. 2013. Generating student feedback from time-series data using reinforcement learning. In *Proc. of 14th European Workshop on Natural Language Generation*, pages 115–124.

Dimitra Gkatzia, Helen Hastie, and Oliver Lemon. 2014. Comparing multi-label classification with reinforcement learning for summarisation of time-series data. In *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1231–1240.

Shinichiro Ishikawa. 2011. *A new horizon in learner corpus studies: The aim of the ICNALE project*, pages 3–11. University of Strathclyde Publishing, Glasgow.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proc. of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606.

Jun'ichi Kakegawa, Hisayuki Kanda, Eitaro Fujioka, Makoto Itami, and Kohji Itoh. 2000. Diagnostic processing of Japanese for computer-assisted second language learning. In *Proc. of 38th Annual Meeting of the Association for Computational Linguistics*, pages 537–546.

Masahiro Kaneko, Yuya Sakaizawa, and Mamoru Komachi. 2017. Grammatical error detection using error- and grammaticality-specific word embeddings. In *Proc. of 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 40–48.

Yi-Huei Lai and Jason Chang. 2019. TellMeWhy: Learning to explain corrective feedback for second language learners. In *Proc. of 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 235–240.

Kathleen F. McCoy, Christopher A. Pennington, and Linda Z. Suri. 1996. English error correction: A syntactic user model based on principled "mal-rule" scoring. In *Proc. of 5th International Conference on User Modeling*, pages 69–66.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proc. of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3197–3206.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating Corpora for Research in Feedback Comment Generation. In *Proc. of the 12th Language Resources and Evaluation Conference*, pages 340–345.

Ryo Nagata, Mikko Vilenius, and Edward Whittaker. 2014. Correcting preposition errors in learner English using error case frames and feedback messages. In *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–764.

Courtney Napoles and Chris Callison-Burch. 2017. Systematically adapting machine translation for grammatical error correction. In *Proc. of 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 345–356.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proc. of Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170.

Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. 2020. A Dataset for Investigating the Impact of Feedback on Student Revision Outcome. In *Proc. of 12th Language Resources and Evaluation Conference*, pages 332–339.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A Simple Recipe for Multilingual Grammatical Error Correction. In *Proc. of 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing*, pages 702–707.