Identifying Feedback Types to Augment Feedback Comment Generation

Maja Stahl and **Henning Wachsmuth**
Leibniz University Hannover, Hannover, Germany
Institute of Artificial Intelligence
{m.stahl, h.wachsmuth}@ai.uni-hannover.de

Abstract

In the context of language learning, feedback comment generation is the task of generating hints or explanatory notes for learner texts that help understand why a part of text is erroneous. This paper presents our approach to the Feedback Comment Generation Shared Task, collocated with the 16th International Natural Language Generation Conference (INLG 2023). The approach augments the generation of feedback comments by a self-supervised identification of feedback types in a multitask-learning setting. Within the shared task, other approaches performed more effective, yet the combined modeling of feedback type classification and feedback comment generation is superior to performing feedback generation only.

1 Introduction

Several studies have dealt with identifying and correcting grammatical errors to help language learners improve their writing skills (Imamura et al., 2012; Bryant et al., 2017; Rozovskaya and Roth, 2019; Grundkiewicz et al., 2019). However, these approaches do not provide learners with a rationale for why a piece of text is erroneous. To help learners better understand and adapt the underlying writing rules, Nagata (2019) introduced the task of *feedback comment generation*: Given a learner text in which some span is known to be erroneous, automatically generate a comment containing helpful hints and explanations. Specifically, the comment should prompt the learner to come up with a solution rather than pointing out an error (grammatical error detection) or correcting it (grammatical error correction).

Towards this end, the Feedback Comment Generation Shared Task (Nagata et al., 2021) at the 16th International Natural Language Generation Conference (INLG 2023) has provided a corpus of erroneous English sentences written by non-native learners of English. Each sentence comes with a feedback comment that is targeted towards a given position of the sentence. The focus is on errors related to the use of prepositions in order to restrict the extensive task of generating feedback to a manageable setting. The generated comments are supposed to explain to the writer why the text part in question is erroneous, possibly with related writing rules. One exemplary instance of the task looks as follows:

Input Text "They can help their father or mother <u>about</u> money that we must use in the university too."

Feedback Comment "«About» is not the appropriate <preposition> to be used when a <noun> follows the structure <help + someone>. Look up the use of the <verb> «help» in a dictionary to learn the appropriate <preposition> to be used."

As our contribution to the shared task, we present an approach that relies on multitask-learning to simultaneously (a) classify the *type* of the target feedback for the given erroneous input sentence and (b) generate an appropriate feedback comment of this type. Since no feedback type labels are given in the data, we tackle the type classification in a self-supervised manner. In particular, we apply an unsupervised clustering based on TF-IDF vector representations of the feedback comments. Each cluster is assumed to represent one feedback type. We then learn a mapping from input texts to feedback types. The rationale is that an explicit distinction between different types of feedback may help to generate targeted feedback comments per type and, hence, more diverse comments for different types. Overall, the generated feedback comments may then better match the input text by exploiting the feedback patterns per comment type.

Our evaluation results in the shared task suggest that the combined modeling of feedback type classification and feedback comment generation is

Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges, pages 31–36
September 11-15, 2023

superior to performing feedback generation only. Our approach improves over sequence-to-sequence baselines in automatic and manual evaluation.

2 Related Work

Supporting non-native speakers of a language to improve their writing skills has been approached from multiple perspectives. So far, however, the main focus has been on detecting and correcting grammatical errors in text.

Early research often targeted only on one common error type, such as incorrect article usage (Han et al., 2006), preposition and determiner usage (Gamon et al., 2008; De Felice and Pulman, 2008), singular and plural usage (Nagata et al., 2006), or false friends (Katrenko, 2012). More recent work proposed approaches to detecting (Nagata et al., 2022) and correcting (Chollampatt et al., 2016; Takahashi et al., 2020; Junczys-Dowmunt et al., 2018) grammatical errors in general using large-scale neural networks, including transformer-based language models. Some works go beyond grammar to assess argumentative structures in learner texts (Wachsmuth et al., 2016; Stab and Gurevych, 2016; Chen et al., 2022). Creutz and Sjöblom (2019) proposed the usefulness of rewriting language learner texts not only to correct errors but also to improve the fluency and naturalness of a text.

The task of feedback comment generation, as proposed by Nagata (2019), goes beyond detecting and correcting errors in that it includes to provide explanations for why some text part is erroneous. With this, language learners can better understand and adapt the underlying writing rules. Hanawa et al. (2021) compared a neural retrieval-based method to a sequence-to-sequence model and a hybrid of these two that edits retrieved feedback comments. They found that the sequence-to-sequence model works best in a setting with few feedback variations, for example, concerning preposition use only. At the same time, the hybrid approach seems most promising for general feedback generation.

3 Task and Data

This section summarizes the Feedback Comment Generation Shared Task as well the data provided as part of the task.

3.1 Task

In the context of the Feedback Comment Generation Shared Task, the definition of feedback comment generation can be summarized as follows (Nagata et al., 2021):[1]

Given an input text and a position known to be erroneous regarding preposition use, automatically generate hints or explanatory notes (feedback comment). The generated feedback comment should explain to the writer why the input text is erroneous at the specified position, possibly with related writing rules. Alternatively, the special token *<NO_COMMENT>* can be generated if an approach cannot generate reliable feedback.

3.2 Data

Each instance in the dataset provided by the organizers consists of an English erroneous input sentence, the position of the error, and a manually written feedback comment targeted towards the error position, as described in Nagata (2019). A total of 4868 training, 170 development, and 215 test instances was provided.

The sentences come from essays of the International Corpus Network of Asian Learners of English (ICNALE) that were written by Asian college students with proficiency levels in English estimated to be between A2 and B2+ in the CEFR metric (Ishikawa, 2013). The essays discuss two topics: (a) "It is important for college students to have a part-time job", and (b) "Smoking should be completely banned at all restaurants in the country". The feedback comments were written by professional annotators with good English skills. They were asked to use special symbols in their writing to highlight specific tokens: $(<, >)$ to surround grammatical terms, $(<<, >>)$ to surround citations from the input text.

4 Approach

We now present our approach to feedback comment generation. Its core idea is to classify the type of feedback to be given and to generate an according feedback comment simultaneously.

4.1 Overview

As illustrated in Figure 1, our approach consists of two main stages:

1. *Feedback Clustering.* We first perform clustering on the TF-IDF vector representation of the training feedback comments in order to identify different feedback types.

[1] `https://fcg.sharedtask.org/task/`, last accessed: 2022-09-12

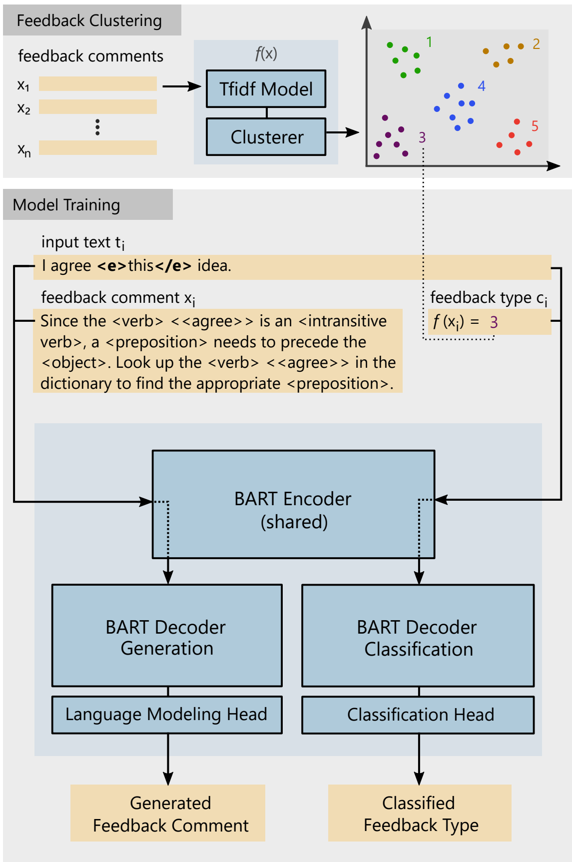

Figure 1: Overview of our approach: First, the training feedback comments are clustered into *feedback types* based on their TF-IDF vector representations. Given an input text and the position of an error, a multitask-learning model then jointly classifies the feedback type and generates the target feedback comment.

2. *Model Training.* Then, a pre-trained language model is trained jointly on feedback type classification and feedback comment generation, using the cluster number from Step 1 as the target label for the classification.

Notice that the feedback type classification is performed based on the erroneous input sentences and not on the target feedback comments, since the latter are not available at inference time. The model is therefore expected to infer the feedback type to be given from the input text only.

4.2 Details

For the feedback clustering, we remove citations from the erroneous input texts as highlighted with $(<<, >>)$ from the feedback comments, to improve the generalizability. For model training and inference, we provide the model with the error position by surrounding the erroneous text part with special tokens $(< e >, < /e >)$, as shown in Figure 1.

For the joint classification and generation, we use a transformer-based encoder-decoder model in a multitask-learning setting. Multitasking is performed by sharing the encoder between the two tasks and combining it with task-specific decoders and language modeling and classification heads, respectively. The training of the model is performed alternately for both tasks, so the encoder weights are updated in each step, while only one decoder and one model head are updated at a time. The hypothesis is that this setting leads to encodings that differ more between different types of feedback comments and are more similar for similar target feedback comments compared to a single task setting. We expect this to help generate more targeted feedback towards the feedback comment types identified in the training data.

5 Evaluation

This section reports on our experiments with joint feedback type classification and feedback comment generation before presenting the evaluation methods and results.

5.1 Experimental setup

In our evaluation, we relied on the following setup:

Feedback Clustering For clustering feedback comments, we use the scikit-learn implementation (Pedregosa et al., 2011) of TF-IDF to transform the training feedback comments into their vector representations. We excluded vocabulary entries with an absolute document frequency below 5 and a relative document frequency above 95% in order to remove rare tokens and stop words. On this basis, we ran k-means clustering with pseudo-random centroid initialization (seed 42). We optimized the number of clusters against the BLEU score (Papineni et al., 2002) of the generated feedback comments and found $k = 12$ clusters to perform best in this regard.

Feedback Type Classification Next, we employed the TF-IDF model and the k-means model to infer feedback types for the validation examples, which we then used to evaluate classification performance. On the validation set, our model achieved a macro-averaged F_1-score of 0.80 for feedback type classification. The score varied between 0.59 and 0.89 for numbers of clusters between 6 and 30.

Approach	Automatic (BLEU)	Manual (F_1)
Generation-BART	0.394	n/a
Generation-Pointer (Nagata et al., 2021)	0.334	0.312
Multitask-BART (our model)	**0.437**	**0.358**

Table 1: Automatic and manual evaluation results: Our model outperforms both baselines in terms of BLEU score, and it also improves over the shared task baseline of Nagata et al. (2021) in the manual evaluation.

Feedback Comment Generation In our language model experiments, we used the HuggingFace implementation (Wolf et al., 2020) of the pretrained BART language model with 139M parameters (Lewis et al., 2020). Together with the cluster number optimization, we tuned the hyperparameters for the training of the model and found a learning rate of 5^{-5}, batch size of 4, 8 training epochs, and length penalty of 1.0 to perform best regarding the feedback comment generation. Below, our model is called *Multitask-BART*.

Baselines We compare the Multitask-BART model against to two baselines:

- *Generation-BART*. A modification of our model, trained only on feedback comment generation.
- *Generation-Pointer*. The baseline model provided by the shared task organizers, which is an encoder-decoder model with a copy mechanism based on a pointer generator network (Nagata et al., 2021).[2]

5.2 Results

Table 1 presents the results of both the automatic and the manual evaluation.

Automatic Evaluation We automatically assessed the feedback comment generation quality of all models on the test set using BLEU score (Papineni et al., 2002), as suggested by the organizers. Among the evaluated approaches, our proposed model achieves the highest BLEU score (0.437), that is, its output has the highest overlap with the human-written reference comments.

Manual Evaluation In addition, our submitted shared task approach was manually evaluated by the organizers, who compared the generated feedback comments to the corresponding reference feedback comments. A generated feedback comment was considered correct when (1) it contains information similar to the reference and (2) it does not contain information irrelevant to the error position. The overall performance was then measured as F_1-score based on the correctness labels (Nagata et al., 2021).

With an F_1-score of 0.358, our model outperforms over the strong baseline based on a pointer generator network (0.312), even though the performance difference is not as big as in the automatic evaluation. Compared to the other submissions to the shared task, our model achieved the sixth place in the automatic evaluation and the seventh place in the manual evaluation.

Error Analysis To obtain insights into the weaknesses of our approach, we finally looked at those feedback comments generated by the model that were flagged as incorrect by the organizers. We found that the main contents of the comments are often correct or somewhat correct, but the important details, which were highlighted in the target feedback comments by brackets, are wrong. For example, a wrong word is cited from the input text, or a word not present in the input is generated as if it was a citation from the input (using the brackets $<<, >>$). The generated grammatical terms (surrounded by $<, >$) are the other common error of our model, which is more problematic as it cannot be identified easily as an error by a language learner. The organizers made the same observations when they assessed our model output.

6 Conclusion

This paper has described our approach to the Feedback Generation Shared task Collocated with the 16th International Natural Language Generation Conference (INLG 2023). The key idea of our approach is to jointly model the classification of feedback types and the generation of feedback comments in order to exploit found patterns per comment type during the generation. Our experiments suggest that the generation quality improves by modeling both tasks together. We also observed open issues, though, that indicate a wrong integration of parts of the input into the generated output. A refined control of the interaction of input and output may resolve such issues in future work.

[2] https://github.com/k-hanawa/fcg_genchal2022_baseline, last access: 2022-09-12

References

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Wei-Fan Chen, Mei-Hua Chen, Garima Mudgal, and Wachsmuth. 2022. Analyzing culture-specific argument structures in learner essays. In *Proceedings of the 9th Workshop on Argument Mining*.

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1901–1911, Austin, Texas. Association for Computational Linguistics.

Mathias Creutz and Eetu Sjöblom. 2019. Toward automatic improvement of language produced by non-native language learners. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 20–30, Turku, Finland. LiU Electronic Press.

Rachele De Felice and Stephen G. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 169–176, Manchester, UK. Coling 2008 Organizing Committee.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Na-Rae Han, Martin Chororow, and Claudia Leacock. 2006. Detecting errors in english article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

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.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 388–392, Jeju Island, Korea. Association for Computational Linguistics.

Shin'ichiro Ishikawa. 2013. The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner corpus studies in Asia and the World*, 1:91–118.

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

Sophia Katrenko. 2012. "could you make me a favour and do coffee, please?": Implications for automatic error correction in English and Dutch. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 49–53, Montréal, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Ryo Nagata, Atsuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of the 21st International Conference* *on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 241–248, Sydney, Australia. Association for Computational Linguistics.

Ryo Nagata, Manabu Kimura, and Kazuaki Hanawa. 2022. Exploring the capacity of a large-scale masked language model to recognize grammatical errors. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4107–4118, Dublin, Ireland. 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.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118. Association for Computational Linguistics.

Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. 2020. Grammatical error correction using pseudo learner corpus considering learner's error tendency. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–32, Online. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.