Sentence-level Feedback Generation for English Language Learners: Does Data Augmentation Help?

Shabnam Behzad
Georgetown University
shabnam@cs.georgetown.edu

Amir Zeldes **Nathan Schneider**
Georgetown University
{amir.zeldes, nathan.schneider}@georgetown.edu

Abstract

In this paper, we present strong baselines for the task of Feedback Comment Generation for Writing Learning. Given a sentence and an error span, the task is to generate a feedback comment explaining the error. Sentences and feedback comments are both in English. We experiment with LLMs and also create multiple pseudo datasets for the task, investigating how it affects the performance of our system. We present our results for the task along with extensive analysis of the generated comments with the aim of aiding future studies in feedback comment generation for English language learners.

1 Introduction

Grammatical error correction has been vastly studied recently in the NLP community (Wang et al., 2021), but it is not always sufficient to merely provide the learner with a correction; in many cases, explicit feedback can facilitate the learning process. Language learners can revise improperly employed linguistic elements by reviewing feedback containing information on the error such as an explanation of why the usage is incorrect and suggestions on how to correct it. This will also help the user avoid making similar errors in the future (Pilan et al., 2020).

In this paper, we focus on preposition errors made by English language learners. Some studies have shown that the majority of syntactic errors made by English language learners are prepositional errors of substitution, omission, and addition (Lorincz and Gordon, 2012). Prepositions are challenging for language learners to master since they are highly frequent; short, unstressed and perceptually weak; and can have several different senses which may not map onto their native languages (Tyler and Evans, 2003; Morimoto and Loewen, 2007; Johansson Falck, 2015).

The task of feedback generation hasn't been explored much until recently when Nagata (2019) proposed the feedback comment generation task and a corpus (Nagata et al., 2020) and then organized the GenChal 2022: FCG (Feedback Comment Generation for Writing Learning) shared task (Nagata et al., 2021). In this task, a system generates an explanation note, given a sentence and a span that indicates the error in the sentence.

Later, Hanawa et al. (2021, 2022) explored different baselines for this task, including a neural-retrieval-based method, a pointer-generator-based seq2seq model, and a retrieve-and-edit method. For preposition-related errors, they found the pointer-generator-based seq2seq model performs the best.

In this paper, we describe our submission to GenChal 2022: FCG (Nagata et al., 2021). We use a simple encoder-decoder model to tackle the task and provide extensive analysis of the different aspects of the task. Our contributions in this paper are as follows:

- We present a simple but strong baseline for the FCG task which is currently ranked third on the leaderboard (team *GU*, BLEU score 0.472; top leaderboard score is 0.486).
- We look into data augmentation techniques and their usefulness for this task.
- We analyze samples that were marked as incorrect by human evaluators and categorize the errors made by our system.
- We further investigate the automatic evaluation metric used for the task and whether or not it is in line with human evaluations.

2 Experiments

2.1 Data

We use data provided by Nagata et al. (2021). The sentences come from essays in ICNALE (The International Corpus Network of Asian Learners of English; Ishikawa, 2013). ICNALE contains essays on two topics: "It is important for college students to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country".

Nagata et al. (2021) hired annotators to annotate a subset of the data for preposition errors. Annotators manually annotated all preposition errors with feedback comments in Japanese (Nagata, 2019) and later translated these comments to English for the FCG shared task. The corpus consists of 4868, 170, and 215 sentences in the train, dev, and test sets respectively. The input for the task is a sentence and a span of the text which contains the error. The output is a string that explains why the span is erroneous. Example:

> *Input:* And we can put posters to remind the smokers the risks they are taking . 37:48
> *Output:* When the <verb> «remind» is used to express " to cause someone to remember something", "someone" is an <object> and a <preposition> needs to precede "something". Look up the use of the <verb> «remind» in a dictionary and add the appropriate <preposition> according to the context.

If a sentence contains more than one preposition error, it can appear more than once in the training set, each time with a different span offset. We incorporate span offsets by adding special characters before and after the erroneous span before encoding the text. For example, the above input sentence becomes: *And we can put posters to remind the *** smokers the *** risks they are taking .* We do not perform any further preprocessing since the text is already processed and tokenized. We used *** as special characters in our setting but the tokenizer behaved the same way when using other characters such as {.

2.2 Experimental Setting

As a baseline for this task, we use T5 (Raffel et al., 2020) as our model. T5 is an encoder-decoder model built on top of the transformer architecture (Vaswani et al., 2017) which is pretrained using a combination of masked language modeling and multitask training such as summarization, machine translation, and sentiment classification.

In our experiments, we encode the essay sentences and fine-tune the model to decode feedback comments. We fine-tune *T5-Large* (770M parameters) with the following hyper-parameters: batch size = 8, learning rate = 0.0001 and maximum training epoch = 50.[1]

2.3 Pseudo Data

We experiment with a few other settings, trying to leverage pseudo data. To create the pseudo data, we select random sentences that are in the same corpus as the gold data (an ICNALE subset that includes correction of sentences; Ishikawa, 2018) but are not included in the FCG shared task train/dev/test sets. Since the focus of the FCG shared task is on preposition errors, we use ERRANT (Felice et al., 2016; Bryant et al., 2017) to annotate error types in these sentences. Then we keep the samples that have preposition-related errors. This gave us 544 additional sentences. To obtain comments for these new sentences, we use our fine-tuned T5 model and generate comments for these samples. We experiment with the pseudo data in two ways:

Multi-stage fine-tuning Fine-tune T5 on pseudo data, and then fine-tune that model on gold training data.

Combined fine-tuning Combine pseudo and gold data, and fine-tune T5 on the combination.

Other than experimenting with pseudo data created from the same learner corpus, we create a large pseudo dataset from other learner corpora, W&I+LOCNESS (Bryant et al., 2019; Granger, 2014). W&I (Write & Improve) is an online web platform in which users from around the world submit letters, stories, articles, and essays, and the system provides automated feedback. Some of these submissions have been further corrected by annotators. LOCNESS consists of essays written by native British and American undergraduates on different topics.

Using ERRANT, we select sentences from W&I+LOCNESS that have preposition errors. This resulted in 6,973 sentences. For the grammatical error correction task, Kiyono et al. (2019) suggests that when the amount of pseudo data and gold data is balanced, concatenating them for training works better (combined fine-tuning), but when the amount of data is unbalanced, a multi-step approach works better (multi-stage fine-tuning). Here, we investigate this by comparing conditions where the pseudo data is limited to 5,000 samples (balanced) versus conditions with all 6,973 pseudo samples (unbalanced).

[1] https://github.com/shabnam-b/GU-FCG-2022

Model	Dev BLEU	Test BLEU	Human Evaluation F1 (Test)
FCG Shared Task Baseline	46.30	33.40	31.16
F/t T5 Large (No pseudo data)	**57.29**	47.11	58.60
Multi-stage f/t (ICNALE)	55.01	46.76	–
Combined f/t (ICNALE)	55.55	**47.25**	**61.90**
Multi-stage f/t (WIL, balanced)	55.46	45.95	–
Combined f/t (WIL, balanced)	57.05	46.91	61.40
Multi-stage f/t (WIL, unbalanced)	55.05	44.97	–
Combined f/t (WIL, unbalanced)	**57.29**	45.36	–

Table 1: Comparison of models on dev and test sets. *WIL* refers W&I+LOCNESS. The gold training data on which T5 is fine-tuned contains 4,868 samples. The multi-stage fine-tuning and combined fine-tuning conditions make use of data augmentation, supplementing the gold training data with pseudo data. The pseudo data consists of 5,000 samples in the balanced setting and 6,973 samples in the unbalanced setting. There are 170 and 215 samples in the dev and test sets, respectively. Best scores in each column are bolded.

3 Results and Analysis

Results of our experiments are available in Table 1. We compared against the official shared task baseline system, which was an encoder-decoder with a copy mechanism based on a pointer generator network.

3.1 Automatic Evaluation

We use test set BLEU scores to compare all the conditions in Table 1. On this metric, all systems based on T5 give improvements of 12+ points over the official baseline. The gain for our best model (which uses pseudo data from ICNALE for combined fine-tuning) is almost 14 points.

Multi-stage vs. combined fine-tuning In all our experiments, *Combined f/t* showed better performance compared to *Multi-stage f/t* (by a difference of 1 BLEU point or less).

Balanced vs. unbalanced In our experimental setup, using a larger pseudo dataset hurt the performance in both *Combined f/t* and *Multi-stage f/t* settings. One possible explanation is the amount of noise that is being introduced to the system by pseudo data. Creating pseudo data with different techniques might show different results.

In-domain vs. out-of-domain pseudo data Even though our in-domain pseudo data was very small (544 sentences), it was more effective than larger amounts of out-of-domain pseudo data. An intuitive explanation for this case is that ICNALE contains essays on only two specific topics: "It is important for college students to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country". Since the FCG shared task test set comes from ICNALE, a more general model fine-tuned on pseudo data from other corpora might not necessarily perform well on this test set. It seems likely that the model trained on multiple datasets would be more robust in realistic settings testing on other domains.

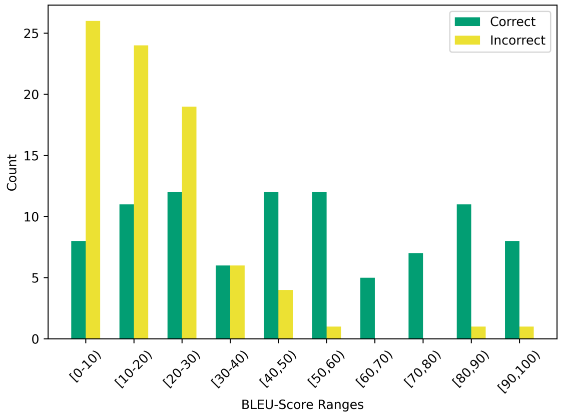

Figure 1: Comparison of human evaluations (correct or incorrect system generated feedback comment) with automatic evaluation metric (BLEU score)

3.2 Human Evaluation

Shared task organizers provided us with the human evaluation of three of our systems (4th column in Table 1). In this evaluation, each system output is compared to the corresponding reference. System output is regarded as appropriate if the following criteria are met: A) it contains information similar to the reference and B) it does not contain information that is irrelevant to the erroneous span. The performance is measured by recall, precision, and F1 based on correct/incorrect outputs.[2]

Based on this human evaluation, our best model achieved an F1 score of 61.09 (this was not our

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

official submission to the shared task, but falls just behind the top leaderboard score[3] of 62.15). Comparing the performance of different systems, human evaluation results and test set BLEU scores seem to be consistent. We investigate this further for our top system, comparing human labels (correct or incorrect) with the BLEU score for each sample in the test set. Results are available in Figure 1. Based on this analysis, when BLEU score is higher than 60%, it is mostly in line with the human evaluations. We also observe that about 49 samples (23% of the test set) are indeed correct, but get a BLEU score below 50. This is due to system-generated comments not having much overlap with the gold feedback comment, despite being correct (Sulem et al., 2018; Nema and Khapra, 2018).

Lastly, we look at 50% of samples where the model-generated comment was labeled as incorrect in human evaluation. We observed that generated comments are very fluent and follow the templates FCG annotators used. In cases where the system output was labeled as incorrect, some of the patterns we observed are as follows:

Completely incorrect comment (≈54%): The model's generated comment includes incorrect suggestions and explanations (first and second example in Table 2). Interestingly, we noticed that the model made the same wrong suggestion in different sentences, containing the same type of error (for example, usage of "on" when it means *sticking to, or hanging from a surface* such as "on the door", "on the wall"). Possible explanations for these cases are that 1) similar errors were not seen during training and 2) in most cases, the sentence contains other errors within the same span or nearby tokens, which presumably makes it hard for the model to understand what the learner was trying to say.

Correct explanation, but incorrect suggestion (≈22%): In these cases, the model gives the right correction, but the explanation is incorrect or incomplete (third example in Table 2).

Correct suggestion, but incorrect evaluation (≈14%): In many cases, the model gives the correct suggestion but the comment starts with something along the lines of "It is not grammatically incorrect to use the ...", even though the usage is indeed incorrect (fourth example in Table 2).

Human annotation errors (≈12%): In a few cases, we believe the system-generated comment is correct, but wrongly labeled as incorrect.

Looking at some positive examples, there are many cases where the model generates completely valid comments. In 19% of cases (41 samples), the model generates exactly the same comment as the reference. In all of these instances, the exact comment was seen during training. There were another 51 comments in the test set that were seen during training, and the model was able to generate a correct comment (but not exactly the same) in 38 cases of those. In many cases, the system output has minor differences compared to the gold output but there are also cases where the generated comment is completely different. Examples appear in Table 3.

4 Discussion

In this paper, we presented a simple baseline for sentence-level feedback generation for English language learners. We investigated the effect of using pseudo data for the task and provided an analysis of system outputs and metrics used.

Does data augmentation help? Incorporating pseudo data gave us very slight BLEU score improvements (0.14) over the T5 model with no pseudo data. But in human evaluations, we observe a 3.3% improvement in F1. Our experiments show that pseudo data could potentially improve results if carefully created and incorporated. Studies on similar tasks such as grammatical error correction have shown that there are many important factors when creating pseudo data such as choice of learner corpora, the method of generating the pseudo data, and error tendency of learners (Kiyono et al., 2019; White and Rozovskaya, 2020; Takahashi et al., 2020). Future studies should further investigate similar important factors for this task and their effectiveness.

Is this a general evaluation of grammatical error feedback capability? Even though our model outperforms the baseline for this task by a large margin, we think it is not a good indicator of current models' performance for the feedback comment generation task in general, due to the following limitations: 1) The topics covered in the train/dev/test sets are very limited and only related to smoking and part-time jobs. Considering the topics and the fact that learners may have a limited vocabulary in the target language, the model is likely performing well because it has seen most of the errors during training, not because it has developed the capacity to recognize and comment on grammat-

[3] As of 14 December 2022

Input Sentence	Gold Output	System Output
Where there is food **on** the smell of cigarette smoke , food soaks Damage to people who go around If you smoke , you know everyone is very often not a good point .	Use the <preposition> ‘with’ instead of «on» to express “food with the smell of cigarettes”.	The <preposition> «on» is not used with ’smell’ as it is meant to indicate a connection. Use the <preposition> ’in’ instead.
The direct means when you smoking , some smoke and chemical access **to** your body after that all bad diseases will show the results .	Since the <verb> «access» is a <transitive verb>, a <preposition> is not necessary.	«To» is not the <preposition> used with the <noun> «access» to indicate a point of arrival. Look up the <noun> «access» in the dictionary and find the appropriate <preposition>.
And , it is the best **of** us to acquire various skills , manners without lazing about study .	<For + noun> should be used as the subject of a <to-infinitive> describing the <formal subject> «it».	«Of» is not the <preposition> used with the <adjective> «best» to indicate the recipient of an action. Think of a <preposition> that indicates destination.
We are as engineer usually work in an industry or **on** a factory , we use our attitude , knowledge to get a job which is comfortable with us .	When the <verb> «work» is followed by the <preposition> «on», it means “to spend time repairing or improving something.” Use the <preposition> “in” or ”at” to indicate where you work.	It is not incorrect to use the <preposition> «on» in this case, but «on» describes the state of being physically inside the factory building. Use in’ instead to refer to business as a situation or an organization”.

Table 2: Examples of incorrect system-generated comments. Error span tokens are bold and in red.

Input Sentence	Gold Output	System Output
Smoking sometimes makes people uncomfortable , so smoking should be banned **at** the public place like the restaurants .	The <preposition> «at» may be used with «place» to indicate a place where something takes place, but it is more common to use ’in’ instead.	The <preposition> «at» may be used with «place» to indicate a place where something takes place, but it is more common to use ’in’ instead.
In the past , most **of** college students probably had no part-time job	A group of something specific follows «most of». Use «most» as an <adjective> rather than a <noun> when simply referring to students in general.	«Most of» is followed by a <pronoun> or a <noun> with a <determiner> such as ’the’, ’these’, or’my’. Otherwise, «most» comes just before the <noun> to be qualified as an <adjective>.
We meet new people and learn about service for customer , how difficult or hard to get money before we **graduate school** .	«Graduate» needs to be followed by a <preposition> when used as an <intransitive verb>.	Since the <verb> «graduate» is an <intransitive verb>, the <object> requires a <preposition>. Look up the use of the <verb> «graduate» in the dictionary.
Parents worry **about** they waste time and hurt themselves .	The <preposition> «about» follows the <verb> «worry» when a <noun> follows. When there is a <clause> in the structure of <S + V>, use a <that-clause>.	The <verb> «worry» is a <transitive verb> that takes a <that-clause> as an <object> which does not require a <preposition>.

Table 3: Examples of correct system-generated comments. Error span tokens are bold and in red.

ical errors in general. 2) The task is focused on preposition-related errors, which makes the errors and comments even more limited than in realistic settings. 3) Most of the comments follow a specific template, which made it easier for the model to learn the patterns. 4) Many reference comments consist of boilerplate—very general suggestions such as *Look up the use of the <verb> «prohibit» in a dictionary and rewrite the sentence using the appropriate structure.* The model also generates many such boilerplate sentences. Of course, a general correct comment is better than a detailed incorrect comment, but we think a model that is able to give more specific suggestions would be of more use to learners.

With these limitations in mind, we still believe this could be a first step toward better and more robust feedback comment generation systems, and we view the organization of the shared task and the release of the data as important milestones for making progress in this research area.

5 Acknowledgements

We thank the shared task organizers and anonymous reviewers for their insightful comments. This research was supported in part by NSF award IIS-2144881.

References

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

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.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Sylviane Granger. 2014. The computer learner corpus: a versatile new source of data for SLA research. In *Learner English on computer*, pages 3–18. Routledge.

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.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2022. Analyzing methods for generating feedback comments for language learners. *Journal of Natural Language Processing*, 29(3):901–924.

Shinichiro Ishikawa. 2018. The ICNALE edited essays: A dataset for analysis of L2 English learner essays based on a new integrative viewpoint. *English Corpus Studies*, 25:117–130.

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(1):91–118.

Marlene Johansson Falck. 2015. Linguistic theory and good practice: How cognitive linguistics could influence the teaching and learning of English prepositions. *Lindgren, Eva, & Janet Enever (ed.), Språkdidaktik: researching language teaching and learning (pp. 61-73). Umeå: Umeå Universitet.*

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. 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 1236–1242, Hong Kong, China. Association for Computational Linguistics.

Kristen Lorincz and Rebekah Gordon. 2012. Difficulties in learning prepositions and possible solutions. *Linguistic Portfolios*, 1(1):14.

Shun Morimoto and Shawn Loewen. 2007. A comparison of the effects of image-schema-based instruction and translation-based instruction on the acquisition of L2 polysemous words. *Language Teaching Research*, 11(3):347–372.

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, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 340–345, Marseille, France. European Language Resources Association.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. 2020. A dataset for investigating the impact of feedback on student revision outcome. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 332–339, Marseille, France. European Language Resources Association.

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.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. 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.

Andrea Tyler and Vyvyan Evans. 2003. *The semantics of English prepositions: Spatial scenes, Embodied meaning, and Cognition*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems (NIPS)*.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–51.

Max White and Alla Rozovskaya. 2020. A comparative study of synthetic data generation methods for grammatical error correction. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 198–208, Seattle, WA, USA → Online. Association for Computational Linguistics.