Abstract

Currently available grammatical error correction (GEC) datasets are compiled using well-formed written text, limiting the applicability of these datasets to other domains such as informal writing and dialog. In this paper, we present a novel parallel GEC dataset drawn from open-domain chatbot conversations; this dataset is, to our knowledge, the first GEC dataset targeted to a conversational setting. To demonstrate the utility of the dataset, we use our annotated data to fine-tune a state-of-the-art GEC model, resulting in a 16 point increase in model precision. This is of particular importance in a GEC model, as model precision is considered more important than recall in GEC tasks since false positives could lead to serious confusion in language learners. We also present a detailed annotation scheme which ranks errors by perceived impact on comprehensibility, making our dataset both reproducible and extensible. Experimental results show the effectiveness of our data in improving GEC model performance in conversational scenario.

1 Introduction

In recent years, both researchers and businesses have attempted to build effective educational chatbots to help language learners improve their conversational skills in a second language (primarily English) (Huang et al., 2021). However, many such systems, such as GenieTutor Plus (Huang et al., 2017), use rule-based dialog engines, and thus do not take advantage of recent developments in dialog generation using Transformer models, which have vastly improved the quality of modern chatbots (Liang et al., 2020). Extant dialog systems for conversational language learning can be broadly classified into two types. In the first type, the chatbot serves as a teacher and repeatedly asks the user questions to test acquisition of specific words, grammar rules, and other pedagogical targets. In the second type, the chatbot serves as a conversational partner, encouraging users to chat with it and, in some cases, providing corrective feedback to learners (Fryer et al., 2020). It is this latter type which we hope to improve using the dataset proposed in the present paper.

Grammatical error correction (GEC) models are needed to generate appropriate corrective feedback for this second type of educational chatbot. However, current GEC datasets all focus on written essays, a domain which differs markedly from conversational speech in both syntax and style. Also, context is an important factor in conversation and could potentially determine the outcome of error corrections. As a result, datasets comprised of written text produce poor results when applied to dialog (Davidson et al., 2019). Unfortunately, no dataset of error-annotated English second language learner dialog currently exists on which researchers can train and evaluate conversational GEC models. In this work we address this lack of data by developing a high-quality, error-annotated dataset of learner dialog collected from an online educational chatbot. To demonstrate the utility of the proposed dataset, we train and evaluate a state-of-the-art (SOTA) GEC model on our newly developed dataset.

2 Related Work

As with many NLP tasks, the current state-of-the-art in grammatical error correction (GEC) involves using large Transformer-based language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). To evaluate the utility of our dataset, we use Omelianchuk et al. (2020)’s GECToR model.

* Authors contributed equally to this work.

1 Data is available at https://github.com/yuanxun-yx/eracond
which reframes GEC as a sequence labelling task rather than a monolingual machine translation task. GECToR achieves SOTA results on the test corpus used for the BEA 2019 Shared Task on Grammatical Error Correction (Bryant et al., 2019). Other promising supervised GEC models are proposed by Stahlberg and Kumar (2021) and Rothe et al. (2021), who achieve SOTA results on the JFLEG (Napoles et al., 2017) and CoNLL-2014 (Ng et al., 2014) GEC datasets, respectively. Both models achieve these results by combining innovative methods of synthetic data generation with large pretrained transformer language models.

Recent work related to the development of datasets for grammatical error correction include Napoles et al. (2019) who present a dataset of native and non-native English writing. Trinh and Rozovskaya (2021) propose a new parallel dataset of Russian student writing. These datasets add to the growing number of GEC datasets available to the research community. However, no GEC dataset that contains conversational data, in English or any other language, is currently available. We seek to begin closing this gap with the present research.

3 Data Collection

3.1 Data Collection Process

We collected 186 dialogs containing 1735 user utterance turns of open-domain dialog data by deploying BlenderBot (Roller et al., 2020) on Amazon Mechanical Turk (AMT) via LEGOEval (Li et al., 2021). The AMT crowdworkers are L2 English speakers of at least intermediate proficiency. The workers are asked to converse with our chatbot for at least 10 turns (a turn is defined as a bot/user utterance pair) either about movies or the COVID-19 pandemic, resulting in a diverse set of utterances in the dataset. Workers interact with the bot using a typed interface (similar to a messaging app), though we plan to expand this to an ASR-driven system in future work.

3.2 Annotation

After collecting open-domain dialog data as described above, we manually revised each user utterance to correct any non-standard or ungrammatical English usage. All dialogs are corrected by two annotators, providing multiple corrected targets for system evaluation. Our goal was to apply the minimum number of edits needed to make the utterance conform to standard written English while remaining as faithful to the source as possible.

Since we are dealing with online chat conversations, our data is more casual than the more formal written data used to train the original GECToR model. Moreover, because our data consists of human-machine conversations involving English language learners of intermediate level, users are assumed to know basic grammatical rules. Therefore, we wanted to give users the flexibility of choosing to limit feedback—such as only receiving feedback on major lexical and syntactic errors. Importantly, suggesting an excessive number of corrections could overwhelm a less proficient user or possibly irritate a more proficient participant, resulting in reduced user enjoyment and engagement (Koltovskaia, 2020). This goal of allowing users to adapt system output to their individual needs is the primary motivation for our tiered organization of corrections presented in Section 3.3.

With these goals in mind, we designed our annotated dataset to conform to the rules of standard written English with two exceptions: internet shorthand and slang, and short responses which are incomplete sentences. We also made fluency edits (Napoles et al., 2017) of semantic and sentence construction errors, particularly those related to lexical choice, omission, and word order. For example, the source line The movie tell about a poor girl that meet a prince and in love for him, suffers from non-native-like word choice. We corrected this utterance to the movie tells about a poor girl that meets a prince and falls in love with him. We made these corrections with the intention of creating ground truth utterances which are as semantically similar to the source as possible.

3.3 Error Types

We organized our annotated corrections into a 3-level structure based on a perceived ranking of how errors impact the ability of interlocutors to understand what the user is saying, as shown in Table 1. As such, we focus primarily on lexical, syntactic and usage errors (Ferris, 2011; Touchie, 1986), while leaving mechanical errors to the lowest-priority category. This 3-level structure is reflected in our modified ERRANT toolkit and M2 format.

For Level 1, our logic is that others are generally still able to understand a message when it is miss-
ing sentence-final punctuation or when a word is not properly capitalized. Because they are of at least intermediate English proficiency, participants can be assumed to know the underlying rules related to punctuation and capitalization; their errors result rather from inattentiveness (Sermsook et al., 2017) and the informal nature of the conversational genre (Cohen and Robbins, 1976). Consider Ex. 1 in Table 2: the syntactic structure of the sentence makes clear that the user’s response is listing names of actors despite the lack of capitalization and sentence-final punctuation.

For Level 2, our logic is that people are likely able to understand a message despite usage of acronyms, abbreviations, non-English internet slang, or a missing apostrophe. The use of these forms in text-based online conversation is to be expected, since these types of abbreviations are becoming more commonplace in all student writing (Purcell et al., 2013; Thangaraj and Maniam, 2015). However, such cases could potentially lead to misunderstanding, especially when conversing with someone of a different generation or sociolinguistic background. Therefore, we categorize these non-standard forms as moderate “errors” (though they are not errors in the traditional sense). We do not consider these non-standard forms as significant because our assumption is that the writer intentionally chose to use these forms for brevity and in the spirit of informality common in online chat (Forsythand and Martell, 2007).

Finally, for Level 3 we include errors which are likely to result misunderstanding or misinterpretation of a message. As we can see in Ex. 7 in Table 2, the user incorrectly uses the term non-broken instead of unbroken, and exploring instead of exploiting. These lexical errors, particularly the latter, are likely to result in misinterpretation of the speaker’s intended meaning. Similarly, the user makes a subject-verb agreement error in Ex. 8 and a verb tense error in Ex. 9. In the former, the user mistakenly uses a plural verb for a singular subject, while in the latter, the user uses a present tense verb when a past tense verb is needed. Because these errors relate to some of the most fundamental rules in English grammar, such errors must be addressed promptly. Thus, we classify errors of these types as “significant” in our annotation scheme.

4 Dataset Statistics

Table 3 reports statistics related to the composition of the ErAcOnD dataset. All statistics are based on user turns; we omit turns generated by our dialog system, as these are not relevant to training a GEC system to provide feedback to users. Additionally, we exclude utterances which include only stop phrases such as “stop” and “goodbye”, since these are intended to terminate the conversation. Error type tags are generated from annotated parallel data automatically with our modified version of ERRANT, and related figures are averaged across multiple annotators.

As shown in Table 3, Level 3 edits account for about 30% of all errors, which supports the necessity of our proposed categorization feature. The error distribution in our dataset is comparable to that of essay-based GEC datasets, according to statistics provided in Bryant et al. (2019), with the exception of spelling and morphological (inflection) errors, which are substantially higher. While the higher rate of spelling errors is unsurprising in a conversation dataset, the difference in morphological errors warrants further investigation.

5 Grammar Error Correction Model

5.1 Model Architecture

We deployed the SOTA model GECToR (Omelianchuk et al., 2020) as the baseline model. It generates a set of edit operations necessary to correct the input text rather than directly outputting corrected text. Pre-trained Transformer encoders are used to improve performance. The complete training process consists of three stages, and higher quality datasets are used in later stages as training progresses. In order to train a conversational version of the original GECToR model, we replace the original non-conversational data with conversational ones starting from any stage based on the cost. For example, if time is not a concern, we could start using synthetic data on dialog datasets and start training from stage one. Considering the limited size we currently have, we only fine-tune the trained models which are available online ² with level 3 edits. Moreover, the turns are fed directly to the model independently, and so context is not considered. This should be improved in future work.

²https://github.com/grammarly/gector#pretrained-models
### Table 1: Categorization of grammatical errors.

| Level | Impact on Meaning | Error Types |
|-------|-------------------|-------------|
| 1     | Trivial           | Punctuation (excl. apostrophe) & Casing |
| 2     | Moderate          | Acronyms, Abbreviations, Non-English Internet Slang, & Apostrophe |
| 3     | Significant       | SV Agreement, Verb Form, Word Confusion, etc. |

Table 2: Examples from ErAConD dataset.

| Example | Role | Message | Error |
|---------|------|---------|-------|
| 1       | USR  | yes, johnny depp, and brad pitt | Punctuation & Casing |
| 2       | BOT  | are you crazy? Kik is kuerig, the one made to be a coffee maker. | Non-English Internet Slang |
| 3       | USR  | I also like SF movies. It makes me think differently. | Acronym |
| 4       | USR  | it's come for more right now? | Abbreviation |
| 5       | USR  | IT SEEMS DRAMATIC. I'll WATCH. | Apostrophe |
| 6       | USR  | She is not on the line now. Maybe it's nighttime there. | Word Confusion |
| 7       | USR  | I'd say you could help Zhou Yu. He's either unable to create a non-broken hit or he's cheating, exploiting low-wage workers. What do you think? | Verb Form |
| 8       | USR  | It just don't work | SV Agreement |
| 9       | USR  | I have a friend from the US. We have a conversation and I don't know the word bangus in English. So it was hard for me to communicate with her. | Verb Form |

Table 3: Overview of ErAConD dataset.

| | | |
|---|---|---|
| Dialogs | 186 |
| User turns | 1735 |
| User sentences (source) | 2454 |
| Word tokens (source) | 24616 |
| Word types | 2860 |
| Error annotations | 2346.5 |
| Level 3 error annotations | 684.5 |
| # of turns per dialog | 9.33 |
| # of sentences per turn (source) | 1.41 |
| # of tokens per turn (source) | 14.19 |
| # of error annotations per turn | 1.35 |
| # of Level 3 error annotations per turn | 0.39 |
| # of Level 3 error annotations per 100 word tokens | 2.78 |

5.3 Result Analysis

Table 4: Performance of GECToR with each setting. Scores are averaged among five trials.

| Setting | TP | FP | FN | Prec | Rec | F0.5 |
|---------|----|----|----|------|-----|------|
| XLNet   | 62.2 | 412.4 | 145.0 | 0.132 | 0.300 | 0.148 |
| FT XLNet | 42.8 | 34.8 | 162.2 | 0.550 | 0.209 | 0.414 |

Table 4 indicates the efficacy of our data in terms of improving the performance of GECToR. The significant increase of $F_{0.5}$ score is mostly contributed by massive decline of false positives. In other words, after learning from the dataset, the model produces far fewer edits, which helps improve the precision greatly.

6 Conclusions and Future Work

Current GEC datasets all focus on formal writing. We provide a high-quality, error-annotated dataset of English second language learner dialog collected from an online educational chatbot. To demonstrate the utility of our dataset, we train and evaluate a SOTA GEC model on the dataset, deploying GECToR as the baseline. This project lays the groundwork for future work on conversational grammatical error correction, which includes adding other dialog domains and incorporating the native languages of users, with the ultimate goal of designing a GEC dialog system that is customizable for second language learners.

7 Ethical Considerations

Collecting these dialogs for our dataset is difficult such that it requires substantial commitment from participants. And so in order to provide...
as large of a dataset as possible, we utilized the services of Amazon Mechanical Turk as previously mentioned. Given ethical concerns in recent years regarding data acquisition through crowdworkers, we verified that the crowdworkers assigned to our tasks were compensated fairly and treated humanely.

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.

Andrew D Cohen and Margaret Robbins. 1976. Toward assessing interlanguage performance: The relationship between selected errors, learners’ characteristics, and learners’ explanations. Language learning, 26(1):45–66.

Sam Davidson, Dian Yu, and Zhou Yu. 2019. Dependency parsing for spoken dialog systems. 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 1513–1519.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Dana Ferris. 2011. Treatment of error in second language student writing. University of Michigan Press.

Eric N Forsythand and Craig H Martell. 2007. Lexical and discourse analysis of online chat dialog. In International Conference on Semantic Computing (ICSC 2007), pages 19–26. IEEE.

Luke Fryer, David Coniam, Rollo Carpenter, and Diana Lăpuşneanu. 2020. Bots for language learning now: Current and future directions.

Jin-Xia Huang, Kyung-Soon Lee, Oh-Woog Kwon, and Young-Kil Kim. 2017. A chatbot for a dialogue-based second language learning system. CALL in a climate of change: adapting to turbulent global conditions—short papers from EUROCALL, pages 151–156.

Weijiao Huang, Khe Foon Hew, and Luke K Fryer. 2021. Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning. Journal of Computer Assisted Learning.

Svetlana Koltovskaia. 2020. Student engagement with automated written corrective feedback (awcf) provided by grammarly: A multiple case study. Assessing Writing, 44:100450.

Yu Li, Josh Arnold, Feifan Yan, Weiyan Shi, and Zhou Yu. 2021. Legoeval: An open-source toolkit for dialogue system evaluation via crowdsourcing. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, et al. 2020. Gunrock 2.0: A user adaptive social conversational system. arXiv preprint arXiv:2011.08906.

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. arXiv preprint arXiv:1907.11692.

Courtney Napoles, Maria Nádejde, and Joel Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. Transactions of the Association for Computational Linguistics, 7:551–566.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. arXiv preprint arXiv:1702.04066.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–14.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chermodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: Tag, not rewrite. arXiv preprint arXiv:2005.12592.

Kristen Purcell, Judy Buchanan, and Linda Friedrich. 2013. The impact of digital tools on student writing and how writing is taught in schools. Washington, DC: Pew Research Center.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. arXiv preprint arXiv:2106.03830.
Alla Rozovskaya and Dan Roth. 2021. How good (really) are grammatical error correction systems? In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2686–2698.

Kanyakorn Sermsook, Jiraporn Liamnimit, and Rattanee kor Pochakorn. 2017. An analysis of errors in written english sentences: A case study of thai efl students. English Language Teaching, 10(3):101–110.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 37–47.

Shalini Raj Thangaraj and Mahendran Maniam. 2015. The influence of netspeak on students’ writing. Journal of Education and Learning, 9(1):45–52.

Hanna Y Touchie. 1986. Second language learning errors: Their types, causes, and treatment. JALT journal, 8(1):75–80.

Viet Anh Trinh and Alla Rozovskaya. 2021. New dataset and strong baselines for the grammatical error correction of russian. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4103–4111.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

A Appendices

A.1 Dataset Statistics and Experiment Results

| Level | Type    | Number | %   |
|-------|---------|--------|-----|
| 1     | PUNCT  | 824.5  | 63.28 |
|       | ORTH   | 478.5  | 36.72 |
|       | Total  | 1303.0 | 55.45 |
| 2     | SPELL  | 0.5    | 0.14 |
|       | PUNCT  | 229.5  | 63.31 |
|       | PREP   | 1.0    | 0.28 |
|       | OTHER  | 124.5  | 34.34 |
|       | NOUN:POSS | 3.5   | 0.97 |
|       | NOUN   | 2.0    | 0.55 |
|       | DET    | 0.5    | 0.14 |
|       | ADJ    | 1.0    | 0.28 |
|       | Total  | 362.5  | 15.43 |
| 3     | SPACE  | 9.5    | 1.39 |
|       | VERB:TENSE | 37.5 | 5.48 |
|       | VERB:SVA | 19.0  | 2.78 |
|       | VERB:INFL | 1.0   | 0.15 |
|       | VERB:FORM | 37.5  | 5.48 |
|       | VERB   | 40.0   | 5.84 |
|       | SPELL  | 115.5  | 16.87 |
|       | SPACE  | 11.0   | 1.61 |
|       | PRON   | 34.0   | 4.97 |
|       | PART   | 69.0   | 10.08 |
|       | OTHER  | 110.0  | 16.07 |
|       | NOUN:POSS | 3.5   | 0.51 |
|       | NOUN:NUM | 35.5  | 5.19 |
|       | NOUN:INFL | 2.5   | 0.37 |
|       | NOUN   | 35.5   | 5.19 |
|       | MORPH  | 28.0   | 4.09 |
|       | DET    | 57.0   | 8.33 |
|       | CONTR  | 4.0    | 0.58 |
|       | CONJ   | 3.5    | 0.51 |
|       | ADV    | 15.0   | 2.19 |
|       | ADJ:FORM | 2.5   | 0.37 |
|       | ADJ    | 9.5    | 1.39 |
|       | Total  | 684.5  | 29.13 |

Table 5: Error type distribution. Type labels were generated using our version of ERRANT.

| Trial No. | Setting | TP | FP | FN | Prec | Rec | F1 |
|-----------|---------|----|----|----|------|-----|----|
| 1         | XLNet   | 61 | 415| 166| 0.128| 0.269| 0.143|
|           | PT XLNet| 47 | 37 | 177| 0.590| 0.210| 0.420|
| 2         | XLNet   | 69 | 358| 145| 0.162| 0.322| 0.240|
|           | PT XLNet| 52 | 39 | 160| 0.571| 0.245| 0.451|
| 3         | XLNet   | 68 | 456| 140| 0.500| 0.181| 0.370|
|           | PT XLNet| 31 | 31 | 140| 0.181| 0.370| 0.260|
| 4         | XLNet   | 63 | 413| 124| 0.132| 0.337| 0.191|
|           | PT XLNet| 42 | 42 | 144| 0.609| 0.231| 0.459|
| 5         | XLNet   | 72 | 428| 164| 0.144| 0.305| 0.184|
|           | PT XLNet| 42 | 42 | 194| 0.512| 0.178| 0.372|

Table 6: Performance of GECToR with each setting in 5 trials.