ErAConD: Error Annotated Conversational Dialog Dataset for Grammatical Error Correction

Anonymous ACL submission

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 conversational dialog. In this paper, we present a novel GEC dataset consisting of parallel original and corrected utterances drawn from open-domain chatbot conversations; this dataset is, to our knowledge, the first GEC dataset targeted to a conversational setting. We also present a detailed annotation scheme which ranks errors by perceived impact on comprehension, making our dataset more representative of real-world language learning applications. To demonstrate the utility of the dataset, we use our annotated data to fine-tune a state-of-the-art GEC model. 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, syntax, 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 we hope to improve using our proposed dataset.

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. As a result, datasets drawn from written sources, such as student essays, produce poor results when applied to dialog (Davidson et al., 2019). Unfortunately, there currently exists no dataset of error-annotated conversational utterances by English second language learners on which researchers can train and evaluate conversational GEC models. In this work we seek to address this lack of data by developing a high-quality, error-annotated dataset of learner dialog collected from an online educational chatbot.[1] To appropriately annotate our data for language learning applications, we introduce a 3-level grammatical error classification structure in order to categorize errors based on severity. Our motivation for this error classification structure is to give users the opportunity to first focus on improving their most serious grammatical errors. To demonstrate the utility of the proposed dataset, we fine-tune and evaluate a state-of-the-art (SOTA) GEC model using 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, which reframes GEC as a sequence labelling task

*Authors contributed equally to this work.

[1]Data will be available on GitHub upon acceptance.

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 include those of 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 combine innovative synthetic data generation methods with large pretrained transformer language models.

Recent work related to the development of datasets for grammatical error correction include Napoles et al. (2019) who presents a dataset of native and non-native English writing. Trinh and Rozovskaya (2021) proposes a new parallel dataset of Russian student writing. These datasets add to the growing number of GEC datasets available to the research community. However, as previously mentioned, 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 who conversed with our bot are L2 English speakers of at least intermediate proficiency. The workers were 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 interacted 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, 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 seen in previous GEC datasets. Moreover, because our data consists of human-machine conversations involving English language learners of intermediate level, users are assumed to know basic English grammar. 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 annotation scheme 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 and syntactically 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.

For Level 1, our logic is that conversational partners are generally still able to understand a message when it is missing sentence-final punctuation or when a word is not properly capital-

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 1: Categorization of grammatical errors.

Example	Message	Error
1	yes, johnny depp, and brad pitt	Punctuation & Casing
2	Ok, what are you talking about? Kkkkkk	Non-English Internet Slang
3	I also like SF movies. It makes me think differently.	Acronym
4	What's your fav movie right now?	Abbreviation
5	IT SEEMS DRAMATIC. ILL WATCH.	Apostrophe
6	She is not on the line now. Maybe its nighttime there.	Apostrophe
7	I'd say you could help Zhou Yu. He's either unable to create a non-broken hit or he's cheating, exploring low-wage workers. What do you think?	Word Confusion
8	It just don't work	SV Agreement
9	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 2: Examples user utterances with error type from ErAConD dataset.

ized. 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 it clear that the user is listing names of actors despite the lack of capitalization and punctuation.

For Level 2, our logic is that interlocutors are likely able to understand a message despite usage of acronyms, abbreviations, non-English internet slang, or a missing apostrophe. An example of such non-English internet slang is shown in Ex. 2 in Table 2. The use of such forms in text-based online conversation is to be expected, since these types of abbreviations are common 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 linguistic 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, we include errors which are likely to result misunderstanding or misinterpretation of a message in Level 3 . 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 treat these errors as "significant" in our annotation scheme.

4 Dataset Statistics

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 tokens	2.78

Table 3: Overview of ErAConD dataset.

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 (i.e. "stop", "goodbye", etc.) since these are intended to terminate the conversation. Our 3-level structure is reflected in our modified ERRANT (Bryant et al., 2019) toolkit and M2 format. Error type tags are generated from annotated parallel data automatically with our modified version of ERRANT[2], and related figures are averaged across multiple annotators. Inspired by Rozovskaya and Roth (2021), our version of ERRANT also enables users to provide grammatically equivalent edits (i.e. changing "I'm" to "I am"), so that ERRANT can recognize them as identical edits.

As shown in Table 3, Level 3 edits account for 29.17% 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 Training process

We aim to train a grammar correction model generates a set of edit operations to correct the input text rather than directly outputting corrected text. It's trained on five datasets, including both synthetic and real data. To train a model that specifically targets to the conversational setting, we fine-tuned the GECToR model[3] proposed by Omelianchuk et al. (2020) on our collected data. The GECToR model was a pre-trained model that finetuned on five datasets involving both synthetic and real written grammar error correction data.

One caveat, we only chose to fine-tune the GECToR model using Level 3 edits in our dataset and ignore the Level 1 and 2 edits, so our model can perform better in conversational settings. Because we want the model to put more focus on critical errors and ignore median and trivial errors where conversational settings can tolerate. In future work, we plan to train all stages of the GECToR model on targeted conversational data. We also plan to integrate conversational context.

Setting	TP	FP	FN	Prec	Rec	$F_{0.5}$
XLNet	72.4	444.6	147.2	0.140	0.330	0.158
FT XLNet	27.1	13.2	191.1	0.683	0.124	0.352

Table 4: Performance of GECToR with each setting. Scores are averaged among 5 runs. Table 6 provides detailed score of every run. XLNet is the baseline GECToR model based on XLNet, and FT XLNet is the fine-tuned GECToR using level 3 edits.

5.2 Result and Analysis

Table 4 indicates the efficacy of our data in terms of improving the performance of the GECToR model. The fine-tuned model outperforms the original in terms of $F_{0.5}$, a metric commonly used in GEC (Omelianchuk et al., 2020). The significant increase in $F_{0.5}$ score results from a massive reduction of false positives. In other words, after we fine-tune GECToR on our dataset, the model produces far fewer edits, which helps improve the precision greatly. 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.

Due to the limited size of the dataset, and the uneven distribution of errors in user utterances, we use 5-fold cross-validation to ensure the reliability of our results. We report the average of five cross-validation runs. One note, we modified ERRANT to allow equivalent edits, our reported results on all models might be slightly higher than original ERRANT-based results.

6 Conclusions and Future Work

We provide the first high-quality, fine-grained error-correction conversation dataset between English second language learner and an educational chatbot. To demonstrate the utility of our dataset, we train and evaluate a SOTA GEC model on the dataset, resulting in a significant improvement in overall model performance for conversational setting. This project lays the groundwork for future work on conversational grammatical error correction (such as adding other dialog domains and incorporating information about the native languages of users) and customized educational dialog system for second language learners.

[2]Code will be available on GitHub upon acceptance.

[3]https://github.com/grammarly/gector#pretrained-models

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.

Besides, the annotators also examined the dataset to make sure it doesn't contain privacy-related or offensive contents.

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 Lăpușneanu. 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 Nădejde, 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 Chernodub, 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 Rattaneekorn 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 Annotation Exceptions

Even though they violate the rules of standard English, we left the following types of errors unchanged in our annotated dataset:

1. *Utterances that are not complete sentences.* For example, response utterances such as *Yes*, *Very good*, and *Me too* are considered correct in our annotation due to their prevalence in informal dialog, although they are not correct in formal writing.
2. *Use of common English internet slang and shorthand expressions.* Slang and shorthand expressions such as *lol* ("laugh out loud") and *u* (short for "you") are not only distinctive to online chat conversations, but also reflective of their casual nature. Additionally, they may be language, culture, and even sub-culture specific. While these terms may not be suitable to a more formal register, they are generally acceptable in the context of informal dialog (Forsythand and Martell, 2007); thus, we do not classify such usage as errors.

A.2 Dataset Statistics

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	WO	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
	PREP	69.0	10.08
	PART	4.0	0.58
	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.

As described in Section 4, Table 5 shows the type distribution of edit type in ErAConD. Type labels were generated using our version of ERRANT, where some of the bugs in the orginal repository were fixed. Levels of edits were first generated by ERRANT, and then manually checked to label Type 2 edits that are hard to be recognized by code (non-English Internet slangs, acronyms and abbreviations). To take all annotators into consideration, the number of was averaged among multiple annotators.

The statistics give us several important insights. First, the number of "significant" errors is slightly higher than in written GEC datasets such as NUCLE. This result shows that grammatical errors are relatively rare in both the conversational and written domain. Additionally, the average length of each sentence is significantly shorter than written GEC datasets. Finally, the error rate data supports our tiered categorization of errors, as the frequency of errors would be much higher than non-conversational datasets if all less significant errors, such as capitalization and punctuation, were included.

A.3 Experimental Results

Table 6 is the full version of Table 4. Some details of experiment are mentioned at Section 5.2. 20% of the dialogs were chosen randomly for the test set and the rest were used for training. Then 5-fold cross-validation was applied and the whole process was run 5 times in total, so as to observe the reliability of our results. We used the recommended parameters of XLNet to train and test GECToR. From the table we can see that the variance of performance among these runs is small. The distribution of Level 3 edits in test and train sets for each run is also represented in Table 7.

Run No.	Setting	TP	FP	FN	Prec	Rec	$F_{0.5}$
1	XLNet	54	395	157	0.120	0.256	0.135
	FT XLNet	21.4	6.8	184.6	0.759	0.104	0.336
2	XLNet	71	506	134	0.123	0.346	0.141
	FT XLNet	24.4	11.0	179.6	0.690	0.120	0.353
3	XLNet	77	437	168	0.150	0.314	0.167
	FT XLNet	25.4	14.6	219.6	0.637	0.104	0.313
4	XLNet	74	404	146	0.155	0.336	0.173
	FT XLNet	22.6	10.4	196.4	0.686	0.103	0.321
5	XLNet	86	481	131	0.152	0.396	0.173
	FT XLNet	41.6	23.2	175.4	0.642	0.192	0.437
Avg.	XLNet	72.4	444.6	147.2	0.140	0.330	0.158
	FT XLNet	27.1	13.2	191.1	0.683	0.124	0.352

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

Type	1		2		3		4		5	
	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
WO	1.03	1.48	1.23	1.42	1.14	1.45	1.74	1.32	0.87	1.49
VERB:TENSE	8.28	4.73	5.33	5.51	5.68	5.43	2.17	6.15	6.11	5.35
VERB:SVA	4.14	2.41	2.46	2.84	2.27	2.90	0.43	3.25	3.06	2.72
VERB:INFL	0.34	0.09	0.00	0.18	0.38	0.09	0.00	0.18	0.00	0.18
VERB:FORM	4.83	5.65	5.33	5.51	4.55	5.70	5.22	5.53	6.55	5.26
VERB	6.21	5.75	4.92	6.04	6.06	5.79	6.09	5.79	4.37	6.14
SPELL	15.17	17.33	17.21	16.80	17.42	16.74	19.13	16.42	18.78	16.49
SPACE	1.38	1.67	1.64	1.60	3.03	1.27	1.30	1.67	0.87	1.75
PRON	8.97	3.89	4.51	5.07	4.92	4.98	3.04	5.36	3.93	5.18
PREP	9.31	10.29	11.07	9.87	10.23	10.05	8.70	10.36	13.97	9.30
PART	1.38	0.37	0.82	0.53	0.38	0.63	0.00	0.70	1.31	0.44
OTHER	16.90	15.85	16.39	16.00	15.53	16.20	17.83	15.72	10.48	17.19
NOUN:POSS	0.00	0.65	0.41	0.53	0.38	0.54	0.87	0.44	0.44	0.53
NOUN:NUM	5.52	5.10	8.61	4.44	5.30	5.16	7.39	4.74	6.55	4.91
NOUN:INFL	0.34	0.37	0.41	0.36	1.52	0.09	0.87	0.26	0.87	0.26
NOUN	1.38	6.21	5.74	5.07	3.41	5.61	4.78	5.27	2.62	5.70
MORPH	2.41	4.54	2.05	4.53	3.79	4.16	5.65	3.78	3.49	4.21
DET	7.59	8.53	7.38	8.53	9.47	8.05	7.83	8.43	11.79	7.63
CONTR	0.00	0.74	0.41	0.62	0.38	0.63	1.30	0.44	0.87	0.53
CONJ	0.34	0.56	0.41	0.53	0.00	0.63	0.87	0.44	0.00	0.61
ADV	2.07	2.22	1.64	2.31	2.65	2.08	2.61	2.11	1.31	2.37
ADJ:FORM	0.34	0.37	0.82	0.27	0.38	0.36	0.43	0.35	0.87	0.26
ADJ	2.07	1.20	1.23	1.42	1.14	1.45	1.74	1.32	0.87	1.49

Table 7: Level 3 error type distribution (%) in train and test sets of 5 runs.