Grammatical Error Correction in Customer Support Chat using Semi Supervision on Pretrained Language models

Nikhilesh Cherukuri[1], Aditya Kiran Brahma[2]
1(Computer Science,IIIT Sri City, AP
Email:nikhilesh.cherukuri001@gmail.com)
2 (Applied Research,Swiggy, KA
Email: adityakiran888@gmail.com)

Abstract:

Understanding the grammatical errors present in the chat conver- sations is crucial in developing chatbot with high quality data. In food delivery platforms, conversational AI development via chatbot platform is continuously built to understand the context of the cus- tomer conversations and suggest the next utterance by retrieving them from a similar scenario occurred in the past. These sugges- tions if used by agents are sometimes manually edited further based on the relevance in current scenario and suggestion quality. The grammatical quality of suggestions play significant role for the agents to utilize conversational AI assistance and provide better customer resolutions in quick and effective manner. In this paper, we analyse a use case of identifying the frequent grammatical er- rors present in the texts typed by the customer care agents and utilize them to build an automatic grammatical error correction model for the data specific to food delivery conversations. We show that using large pretrained encoder-decoder transformer models and systematic fine-tuning with a smaller downstream task specific data (Grammatical error correction) achieved an overall gain of 15.5 % GLEU score compared to the the baseline approach of using the pretrained model alone.

Keywords — Grammatical error correction(GEC), Language models, Transfer learning, Natural Language Understanding(NLU)

I. INTRODUCTION

In the current era of Conversational AI, NLU is extensively used to support the customer care agent's chat conversations. In the food delivery related platform, AI assists the agents by understanding the context of customer chat and matches a similar scenario oc- curred in the past to provide text suggestions in real time. These text suggestions can be utilized and edited by agents before sending to the customer for quick and effective resolutions. As the type of queries and variations of customer sentences are highly diverse, the corpus from which these suggestions are retrieved by the model is huge and contains grammatical mistakes made by the correspond- ing agents in the past. Thus, it requires to build a Grammatical Error Correction (GEC) module to improve the overall grammat- ical quality of suggestions. The main contributions of this paper are: (i) We use an off-the-shelf Encoder-Decoder based model and fine-tune it on food delivery chat conversations by introducing the grammatical errors synthetically (ii) We identify real time gram- matical edits being made by the customer care agents and further finetune the model using this data All the experiments, training and evaluation are done on real time chat datasets. At the end of experiments, we show our findings on this methodology for our task specific data and summarize the solution. We then perform a single-shot evaluation of their performance on the test set.

II. LITERATURE SURVEY

Recently, leveraging pretrained transformer based language models and task specific fine-tuning has achieved advanced state of the art results in natural language understanding tasks.[Kaneko et al. 2020] prove the suitability of this approach by systematically incor- porating masked language models to encoder-decoder architecture and in the experimental results they show an advantage in model performance by maximizing the approximate knowledge gain of the model when querying from the pool of unlabeled data.[Kiyono et al. 2019] prove that systematic incorporation of pseudo data into corpus to train any Encoder-Decoder models yield significantly better results without the need of an external annotation for large parts of the data.Approaches like [Awasthi et al. 2019] predicts in-place edits instead of generating sentences by using a carefully designed edit space, the model iteratively refines its own predic- tions. A similar ideology is followed in [Omelianchuk et al. 2020] where they tackle the GEC problem as a sequence tagging problem instead of a sequence generation task thereby removing decoder from the loop which allows us to parallelize the inference so it runs faster and achieved state of the art results. But it requires enormous complex data annotation very different from the commonly used sequence generation based GEC approaches.Machine translation systems are similar to GEC but the difference lies where the input sentence is changed only for the few erroneous words, characters and punctuation of the source sentence, [Zhao et al. 2019] used this fact and enhanced the current neural architecture by enabling it to copy the unchanged words and the out-of-vocabulary words directly from the source sentence. [Raffel et al. 2019] used an ideol- ogy where they leveraged a unified approach to transfer learning where the objective is to treat every text processing problem as a “text-to-text” problem, i.e input and output of the model are texts. This framework allows the user to utilize the model and make modifications in the objective function and decoding process based on specific down- stream fine-tuning tasks. By utilizing the T5 language model as baseline[Xue et al. 2020], [Rothe et al. 2021] propose a similar frame- work and build Grammatical Error Correction (GEC) model gT5. They make modifications in objectives used by T5 to make it suitable for GEC tasks and perform GEC fine-tuning on CLANG-8 dataset which is a synthetic corpus created by automatically corrupting grammatically correct sentences thereby creating non-grammatical sentences. Our paper extends the idea of leveraging the T5 model and fine-tuning it on food delivery related chat conversations to build the GEC module on customer care agent texts.

III. METHODOLOGY

There are three major steps in our methodology:

Step-1: Choosing a baseline language model

We use transfer learning principle in our implementation where we choose an off-the-shelf

Encoder-Decoder based model which is pre- trained on large amounts of data and make use of the model's generic natural language understanding. We tried several pretrained mod- els [Omelianchuk et al. 2020],[Kaneko et al. 2020],[Chollampatt and Ng 2018],[Rothe et al. 2021](gT5) among them gT5(discussed in Section 2) outperformed on our dataset with a GLEU score of 0.80 on a test dataset verified internally. This score was not sufficiently high and many errors were not getting identified by the model as the texts we use are specific to Indian food delivery orders which may not be present in the large pretrained corpus of the model. In order to enhance the model performance we propose two stage fine-tuning under different settings over the baseline approach. Step-2: Fine-tuning with synthetic data - Stage 1

In order to improve the model and correct more errors specific to the domain, we have to create data in terms of erroneous - correct sen- tences in this domain. So we extract the chat messages from agents having high positive feedback with an assumption that the number of grammar violations in their texts are minimal. We used synthetic data generation proposed and used by [Awasthi et al. 2019] where we randomly introduce grammatical errors into them which are commonly uttered. The synthetic error space introduced in the dataset consists of duplications, deletions,appends,replacements and tense transformations of words and characters wherever applicable. Append,replace and delete operations mostly consist of prepositions,conjunctions,punctuations,articles,pronouns and verbs. Transformations perform inflections like adding suffix (eg: s, d, es, ing, ed) or replacing suffix (eg: s to ing, d to s). These transfor- mations are chosen out of common replaces in the generic training data such that replaced edits map to only few transformation edits. We utilized around 8,40,000 input - output (input is synthetically induced error sentence and output is the ground truth) text pairs of real chat messages for fine-tuning the model.

Step 3: Fine-tuning with real time edits - Stage 2

With the fine-tuning stage 1, we tackle the problem of out of vo- cabulary words by training the baseline model with the prepared food delivery specific dataset and achieved a GLEU score of 0.89. To further improve our model we include the manually identified real time grammatical edits that customer care chat agents are making to the chatbot's suggestions. We identified 618 grammatical edits being made by the agents and expanded the corpus to 3500 samples by searching for the variants in a larger corpus. Upon fine-tuning the previous version of the model on this unique data we achieved a GLEU score of 0.95 on the test set which was around 6.7% improve- ment from previous version. Though the corpus was not large, with the quality of new data we were able to correct sentences which weren't being identified in previous versions.

IV. RESULTS

In Evaluation of GEC models are done using GLEU score [Napoles et al. 2015] and F1 score [Sokolova et al. 2006]. GLEU score is a simple variant of BLEU [Papineni et al. 2002], modified to account for both the source and the reference sentences, and show that it relates much more closely to human judgments. Comparison of various models and their metrics on the test set are reported in Table 1 . gT5 model after two stages of fine-tuning achieved a GLEU score of 0.95 which outperformed all the other models. In Table 2 we show some of the sample corrections made by our best performing model and whether the previous versions of the model were able to correct the message or not (Original sentences are trimmed to accommodate only errors and their contexts).much more closely to human judgments. Comparison of various models and their metrics on the test set are reported in Table 1 . gT5 model after two stages of fine-tuning achieved a GLEU score of 0.95 which outperformed all the other models. In Table 2 we show some of the sample corrections made by our best performing model and whether the previous versions of the model were able to correct the message or not (Original sentences are trimmed to accommodate only errors and their contexts).

GEC Model	**GLEU**	**F1**

gT5 model without fine tuning	0.803	0.817
gT5 model with 1st stage fine-tuning	0.891	0.865
gT5 model with 2nd stage fine-tuning	**0.958**	**0.943**
[Kaneko et al. 2020] model	0.713	0.652
[Chollampatt and Ng 2018] model	0.685	0.651

Table 1: Comparision of different models on GLEU and F1 score

Input	**Output**	**M1**	**M2**	**M3**
the order will deliver to you	the order will be delivered to you	✓	✓	✓
i have process the coupon food50 successfully	i have processed the coupon food50 successfully	×	✓	✓
accept my sincere apology for the inconvenience	accept my × sincere apologies for the inconvenience	×	×	✓
you will get an update an update regarding this	you will get an update and an update regarding this	×	×	×

Table 2: Sample message corrections for gT5 (M1), gT5 after 1st stage finetuning (M2) and gT5 after 2nd stage finetuning

V. CONCLUSION

In this work we show that using an off-the-shelf large Pretrained language model and systematically fine-tuning them to a specific downstream task (GEC) for food delivery related chats related domains provide significantly improved results. Our best model achieves a GLUE score of 0.95,which is a 15% increase from the baseline model. Currently the model is developed for specific types of chat data where customers are enquiring about the whereabouts of the food delivery order. In future we would like to extend our work to other types of order queries.

REFERENCES

1. *Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Pi- ratla. 2019. "Parallel Iterative Edit Models for Local Sequence Transduction." CoRR, abs/1910.02893. http://arxiv.org/abs/1910.02893 arXiv: 1910.02893.*
2. *Christopher Bryant, Mariano Felice, and Ted Briscoe. July 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). Association for Computational Linguistics, Vancouver, Canada, (July 2017), 793–805. doi: 10.18653/v1/P17-1074.*
3. *Shamil Chollampatt and Hwee Tou Ng. 2018. "A Multilayer Convolutional Encoder- Decoder Neural Network for Grammatical Error Correction." CoRR, abs/1801.08831. http://arxiv.org/abs/1801.08831 arXiv: 1801.08831.*
4. *Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. Dec. 2020. "Fine-tuning BERT for Low-Resource Natural Language Understanding via Active Learning." In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), (Dec. 2020), 1158–1171. doi: 10.18653/v1/2020.coling-main.100.*
5. *Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2018. "Near Human-Level Perfor- mance in Grammatical Error Correction with Hybrid Machine Translation." CoRR, abs/1804.05945. http://arxiv.org/abs/1804.05945 arXiv: 1804.05945.*
6. *Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. June 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). Association for Computational Linguistics, New Orleans, Louisiana, (June 2018), 595–606. doi: 10.18653/v1/N18-1055.

7. *Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. "Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction." CoRR, abs/2005.00987. https://arxiv.org/abs/200 5.00987 arXiv: 2005.00987.*

8. *Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. "An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction." CoRR, abs/1909.00502. http://arxiv.org/abs/1909.00502 arXiv: 1909.00502.*

9. *Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. July 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. Association for Computational Linguistics, Online, (July 2020), 7871–7880. doi: 10.18653/v1/2020.acl-main.703.*

10. *Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. June 2019. "Corpora Generation for Grammatical Error Correction." 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). Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 3291–3301. doi: 10.18653/v1/N19-1333.*

11. *Zhenghao Liu, Xiaoyuan Yi, Maosong Sun, Liner Yang, and Tat-Seng Chua. 2021. "Neu- ral Quality Estimation with Multiple Hypotheses for Grammatical Error Correction." CoRR, abs/2105.04443. https://arxiv.org/abs/2105.04443 arXiv: 2105.04443.*

12. *Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. July 2015. "Ground Truth for Grammatical Error Correction Metrics." In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Asso- ciation for Computational Linguistics, Beijing, China, (July 2015), 588–593. doi: 10.3115/v1/P15- 2097.*

13. *Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem N. Chernodub, and Oleksandr Skurzhanskyi. 2020. "GECToR - Grammatical Error Correction: Tag, Not Rewrite." CoRR, abs/2005.12592. https://arxiv.org/abs/2005.12592 arXiv: 2005.12592.*

14. *Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Oct. 2002. "BLEU: a Method for Automatic Evaluation of Machine Translation," (Oct. 2002). doi: 10.3115/1073083.1073135.*

15. *Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." CoRR, abs/1910.10683. http://ar xiv.org/abs/1910.10683 arXiv: 1910.10683.*

16. *Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. "A Simple Recipe for Multilingual Grammatical Error Correction." CoRR, abs/2106.03830. https://arxiv.org/abs/2106.03830 arXiv: 2106.03830.*

17. *Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Jan. 2006. "Beyond Ac- curacy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation." In: vol. Vol. 4304. (Jan. 2006), 1015–1021. isbn: 978-3-540-49787-5. doi: 10.1007/11941439_114.*

18. *Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. "mT5: A massively multilingual pre-trained text-to-text transformer." CoRR, abs/2010.11934. https://arxiv.org/abs/2010.11934 arXiv: 2010.11934.*

19. *Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. "Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data." CoRR, abs/1903.00138. http://arxiv.org/abs/1903.00138 arXiv: 1903.00138.*