Building Proficiency-Adaptive Chatbots for English Language Learners

Lucas Sha[1], Siyan Sylvia Li[2], Xuanming Zhang[2], Zhou Yu[2]
[1]Department of Computer Science, Brown University, Providence, RI, USA
[2]Department of Computer Science, Columbia University, New York, NY, USA

Introduction:

Chatbots powered by LLMs are being used by English learners to practice their language skills in natural, conversational contexts. However, state of the art bots often generate outputs requiring an already high degree of proficiency in English to understand, making it difficult for less proficient English learners to converse with them.

In this project, we explore techniques to make chatbots automatically adapt their outputs in response to the detected proficiency level of the user conversing with them.

Methods:

Our user/chatbot interaction loop relies on the chatbot being able to extract and detect information about the user's language proficiency through proxy measures. We use two proxy measures of the user's proficiency:

(1) Grammatical coherence
(2) Complexity of their language use

For (1) we evaluate a pre-trained model capable of grammatical error detection[1] on the user's input. For (2) we fine-tune an encoder only sentence transformer model on a Kaggle Dataset labeling various sentences and passages with CEFR levels[2], and likewise evaluate it on the user's input.

We use a weighted combination of these two model outputs as an estimate of the users' overall CEFR proficiency, and thus the chatbot's desired output complexity. We have implemented a custom decoding strategy which alters the logits of tokens depending on their closeness to the desired level of complexity.

We use several instruction-tuned LLMs including GPT-2 and Llama-3 models of various sizes to test our chatbot capabilities and final pipeline.

---

[1] Siyan, L., Shao, T., Yu, Z., & Hirschberg, J. (2024, June 25). *EDEN: Empathetic Dialogues for English learning*. arXiv.org. https://arxiv.org/abs/2406.17982

[2] https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts

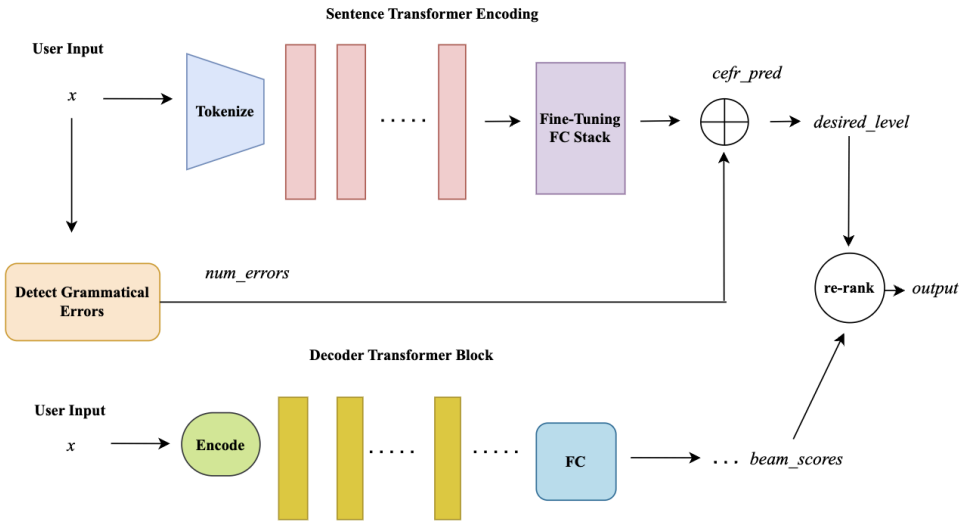

Figure 1: Overview of our proposed method for controlling generation of tokens to align with desired complexity level

Results:
We have not fully constructed the pipeline, nor tested with human input yet.

We have tested the effects of altering desired level on generation; a small table of prompts, desired levels, and resulting outputs is given below:

Prompt :	*'My cute dog'*	*'Historically, America*	*'Despite objections to his policies, the President'*
A1	'My cute dog is one of the best dogs in the world, and I love her so much that she is the only one who loves me.	'Historically, America has had the largest population of immigrants to the U.S. since the Civil War. Today, the country is the second largest in the world.'	'Despite objections to his policies, the President, who took over at the beginning on April 15, said he was trying to change our foreign policy- "I'm sorry"'
C2	'My cute dog is my favorite! I love it! It is a cute little toy and I am very happy!!	'Historically, America has relied on a large-based, well-organized, multi-state government, and it remains one.'	'Despite objections to his policies, the President has said that the U.S. is the only country in the world to have the right to free the press, and that he will not allow the media to interfere with the election of the president.

Model	GPT2	DistilGPT2	DistilGPT2

Our fine-tuned sentence-transformer's performance is not yet satisfactory; generally reaching about 55% validation accuracy on predicting CEFR levels (of which there are 6).

Conclusions:

The preliminary results from shifting the desired level of complexity are interesting. We have not tested on instruction-tuned models yet, but have simply observed its effects on autocomplete style pre-trained only models.

Clearly, the desired level shift does impact the logits and thus the resulting sequences generated. It is not clear that the A1 sequence is always substantially simpler than the C2 sequence, though this also need not be the case—much C1 and C2 vocabulary is extremely rare in everyday use. Some cases such as the 'Historically, America has' prompt do seem to exhibit a clear difference in complexity between the two extremes of generation, however.

References:

Siyan, L., Shao, T., Yu, Z., & Hirschberg, J. (2024, June 25). *EDEN: Empathetic Dialogues for English learning*. arXiv.org. https://arxiv.org/abs/2406.17982

Tyen, G., Brenchley, M., Caines, A., & Buttery, P. (2022). Towards an open-domain chatbot for language practice. *Association for Computational Linguistics*. https://doi.org/10.18653/v1/2022.bea-1.28

Qian, K., Shea, R., Li, Y., Fryer, L. K., & Yu, Z. (2023, April 11). *User Adaptive Language Learning Chatbots with a Curriculum*. arXiv.org. https://arxiv.org/abs/2304.05489

Lu, X., West, P., Zellers, R., Bras, R. L., Bhagavatula, C., & Choi, Y. (2020, October 24). *NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints*. arXiv.org. https://arxiv.org/abs/2010.12884

Chen, Y., Yu, Z., & Hirschberg, J. (2023, August 24). *MULTIPA: a multi-task speech pronunciation assessment model for open response scenarios*. arXiv.org. https://arxiv.org/abs/2308.12490