English Language Learning Chatbot

Ulie Xu[†], **Rakshitha Sankineni**[‡], **Neuman Alkhalil** [‡]
Georgetown University
[†]Department of Linguistics, [‡]Department of Computer Science
{lx59, rs2140, na892}@georgetown.edu

Abstract

In this paper, we present a English language learning chatbot that can converse via texts with beginner-level and advanced-level English learners. The chatbot is fine-tuned on the DialoGPT model with beginner and advanced English dialogue data, and is able to switch the model in between talking to the user after analyzing the user's English language proficiency level. We evaluate our models with both automatic and manual metrics, and find that the chatbot is able to generate more advanced responses with the advanced model and/or given advanced prompts. Furthermore, we also find that the length of prompts also have an effect on the model's language generation.

1 Introduction

With a growing number of English-as-a-Foreign-Language (EFL) students all over the world, the demand for affordable and accessible English learning tools has increased exponentially. There are a variety of different ways through which students can choose to learn English: traditional classroom settings at school, language tutors online or in person, free or subscription-based language practice mobile applications and websites, language immersion by being in a local environment... However, research done by Horwitz et al. (1986) shows that language learners experience anxiety in foreign language classrooms because they fear not having thorough comprehension to all inputs in the target language. In addition, learners who are less advanced in the target language are afraid of other learners' judgment to them making mistakes in the target language. In addition, Canale and Swain (1980) suggest that a communicative approach to second language pedagogy as opposed to a grammatical approach emphasizes that the learner achieves communicative competence, meaning that the learner knows how to use it to the extent that they are able to demonstrate this knowledge in a meaningful communicative situation.

With the applied linguistic ideologies as foundation, we present our English language learning chatbot. Through talking to an English-speaking chatbot, learners will be able to engage in casual English conversations in any domain and learn daily English usage through the communicative approach. In addition, talking to a chatbot in a private one-on-one setting alleviates learning anxiety from a traditional classroom setting. The nature of a chatbot being able to be accessed anytime, anywhere, and for free makes language learning much more affordable and accessible.

Our chatbot was trained on DialoGPT with two datasets, a beginner-level English dataset and an advanced-level English dataset. The dialogue system initially interacts with users with its advanced-level model; after it converses with the user for several turns, it evaluates the user's English proficiency level. If the user is evaluated to be an advanced English learner, the model does not switch and continues chatting with the user; whereas if the user is evaluated to be a beginner English learner, the model switches to a beginner English model, chatting with the user in simpler English.

2 Related Work

The use of chatbots has grown significantly during the past few decades. The first chatbot ever developed in the history of computer science was called ELIZA by Weizenbaum (1966). ELIZA works by identifying key words in minimal context of users' prompts, transforming the key words, and ultimately generating responses to the prompts. The chatbot industry quickly moved on to more advanced conversational AI, with the goal of developing a personal assistant that can converse with users and performing tasks through smart devices such as Amazon Alexa and Apple Siri. With the rise of deep learning from 2015, lots of rule-based chatbot systems have been replaced by end-to-end trainable neural networks (Csaky, 2019). The current state of the art chatbot systems are generally divided into two methodologies: the first is a retrieval method to rank the best response, under the assumption that the pre-constructed conversation datasets already include the proper response to the user's question. The second methodology is a generative method to output the best response. The sequence-to-sequence model (seq2seq) by Sutskever et al. (2014) has been widely adopted for the generative method, which uses recurrent neural networks (RNNs) as the encoder and decoder to convert source sentences into target sentences.

Although there are many open-domain and task-oriented chatbots, there are very limited ones that are built for language learning and are customizable and personalized according to the user's language proficiency. In Huang et al.'s (2022) evaluation of language learning chatbots, they confirmed the positive effects of utilizing chatbots to assist with students' learning as a valid pedagogical approach. In Bibauw et al.'s (2022) meta-analysis of effectiveness studies on dialogue-based computer-assisted language learning (CALL), they further confirmed the effectiveness of form-focused and goal-oriented dialogue systems, as well as system-guided interactions, especially for lower proficiency learners. Shi et al. (2020) has also implemented a transfer-learning-based English learning chatbot by fine-tuning the GPT-2 model with dictionary and ontology graphs as their datasets. They focused on three levels of systematic English learning: phonetics, semantics, and the simulation of free-style conversation. Lastly, there are also an abundance of language learning platforms (LLPs) in the market such as Duolingo, Resetta Stone, Memrise, LingQ, and Busuu. Karasimos (2022) found that although learners of these LLPs were pleased with learning new vocabulary and pronunciation with the gamification feature, some users found a lot of them overfocused on vocabulary. As a result, a communicative approach with less emphasis on vocabulary and grammar, and more on crafting the overall language flow prevails in our chatbot.

3 DialoGPT Model

We selected DialoGPT to fine-tune our chatbot models because of its better performance than many baseline models, relevancy, contentfulness, and context-consistency (Zhang et al., 2019). DialoGPT is a large-scale pre-trained dialogue response generation model for multi-turn conversations. The model is trained on 147M multi-turn dialogues from Reddit discussion threads. Zhang et al. (2019) concluded that sentences generated by DialoGPT are diverse and contain information specific to the source prompt, analogous to the outputs that GPT-2 generates. The model architecture is based on GPT-2, containing a 12-to-48 layer transformer with layer normalization, leveraging a stack of masked multi-head self-attention layers, and using byte pair encodings for tokenization. DialoGPT has three different sizes of pre-trained models: small(117M), medium(345M), and large(762M). Validation loss was used to choose the learning rate, and each model was trained until the validation loss stopped increasing. The small and medium models were trained for up to 5 epochs, while the large model was trained for up to 3 epochs.

4 Data

Two datasets are required for this task: including one with lower-level English and the other with higher-level English, as stated in section 1. Higher-level English data is much easier to obtain because it is relatively closer to the majority of English dialogue data, while lower-level English data was much harder to find, because we require lower-level English to contain less complex language. Therefore, we looked for datasets that, ideally, focus on second language learning data for English learners. However, there were very limited datasets that are constructed in a dialogue format. We came across the Parent-Child Interaction Therapy dataset and the Dailydialog dataset.

4.1 Parent-Child Interaction Therapy Dataset

We first used the Parent-Child Interaction Therapy Dataset to train our beginner model (Huber et al., 2019). The dataset contains 6022 lines of interactions between parents and their children of an average of 4.5 years old, and annotation of parents' dialogue acts by experts. We hypothesized that this dataset would be appropriate to train our beginner model because parents often speak to their children in simple English, and children with an average of 4.5 years old speak in complete but simple English. After we used it to train our beginner model, a sample testing interaction with the beginner model is attached below:

User: *Hello, who are you?*

Beginner Bot: *I made a tower.*
User: *Awe you are so cute!*
Beginner Bot: *This is a big block.*

As indicated by the testing interaction, the chatbot's responses are quite childish and uncontentful. However, our chatbot should ideally be able to converse in simple English rather than child English. We reasoned that the content of the parent-child interaction might be way too simple and monotonous for our task. Moreover, the data is only 283KB in size, which might be too small for the beginner model to train on. As a result, we made the decision to look for a new beginner dataset.

4.2 DailyDialog Dataset

DailyDialog is a collection of real-world conversations that captures how people communicate every day (Li et al., 2017). It is high quality data since it was manually constructed using human-written language, which makes it less noisy than the Reddit Data on which DialoGPT was trained. This dataset contains 13,118 multi-turn dialogues with an average speaker turn of 8 and an average token count of 15 per utterance. In order to bifurcate the data into Beginner English and Advanced English, we decided to calculate the English level of the dialogues with readability scores. We attempted to score using several readability criteria, such as the Readability Consensus [1], but the result was heavily skewed. Then, we applied the textstat[2] library's Dale-Chall Readability Statistic. This metric analyzes the utterance and calculates a score between 0 and 10.0 based on how frequently the 3000 most common vocabulary occur in the utterance. It worked better for scoring our data because of the even split. As a result, we separated the entire Dailydialog dataset into two parts, with any conversations having the Dale-Chall Readability score less or equal to 6.99 as part of the beginner dataset and over 6.99 as part of the advanced dataset.

5 System overview

The data, which originally was in the form of conversations with various utterances, was transformed before fine-tuning into a format that the model could interpret. Seven different contexts were assigned to each utterance so that the chatbot could respond with more relevant responses. We divided the training and test sets at a ratio of 9:1. Our final fine-tuned models' adopted the hyperparameters of 4 epochs, batch size of 4, block size of 512, and a learning rate of $5 \times e^{-5}$.

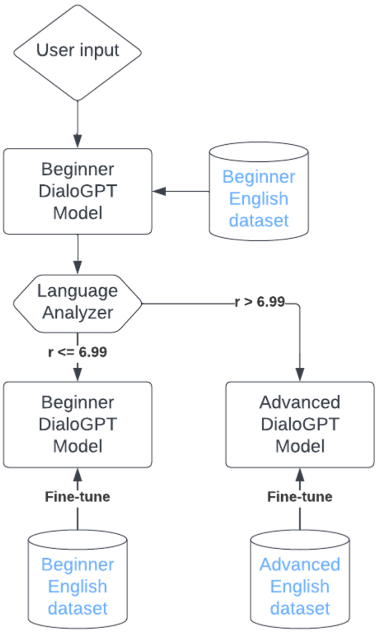

Figure 1: Overall architecture of the English language learning chatbot.

We also experimented with varying hyperparameters for training the model on Google Colaboratory[3]. For both the beginner English model and the advanced English model, we initially trained using 2 epochs and Small-DialoGPT with their respective datasets. Then, for better outcomes and improved chatbot responses, we trained both models using 4 epochs and Small-DialoGPT. Unfortunately, due to lack of computing power, we were unable to train using Large-DialoGPT.

As shown in figure 1, when an user inputs prompts into our chatbot, it first runs through the beginner English model, which was trained on beginner English data, responding to the user in simple English. The reasoning behind the decision to first converse with the user using the beginner model rather than the advanced model is that an advanced user can communicate well with both beginners and advanced models; however, a novice English learner might not be able to understand or communicate with an advanced model, which might discourage the users to keep using the chatbot. After 4 turns of user interaction, the chatbot

[1]Returns average score based on SMOG, Gunning FOG, Flesch-Kincaid, Dale-Chall and other metrics.
[2]https://pypi.org/project/textstat/
[3]https://colab.research.google.com/

evaluates the user's proficiency of the English language with the language analyzer module consisting of the Dale-Chall readability score. If user's readability score is less or equal to 6.99, the chatbot doesn't switch and continues chatting using beginner English model, and if the score is greater than 6.99, the chatbot switches to advanced English model. Refer to appendices A and B for sample transcripts of model switching.

6 Experiments

6.1 Pre-constructed testing prompts

To allow for consistent testing in online evaluation, we manually constructed 70 prompts to feed into the models. The prompts were split into 2 sets of varying lengths: the 50 word-length prompts are short testing prompts, and the 200 word-length prompts are long testing prompts. As a result, we constructed 11 short prompts and 59 long prompts. Of the shorter prompts, 4 of them were scored as advanced level, while the other 7 were scored as beginner level, calculated using the Dale-Chall score. We spent much longer time constructing the 200-word-length prompts; as a result, the long prompts had a much more equal split with 31 beginner level prompts and 28 advanced level prompts. We also constructed the prompts under the premise that having a question at the end of the prompt results in better responses by the chatbot, as the dataset used for fine-tuning the models was mostly in the format of question and answer adjacency pairs. However, this also led to the model not initiating questions as much as a human normally would during an active conversation. The prompts covered a large variety of domains, including movies, sports, food, games, education, etc.

6.2 Testing methods and data cleanup

Testing both the beginner and advanced models was done using the Dale-Chall score previously mentioned. All 70 prompts were used to query the models and their responses were stored in a CSV file along with the original prompt. This CSV file was then parsed using a script which calculated the scores for each response. This data was used in offline evaluation to calculate statistics such as the averages of both the beginner model and the advanced model, as well as calculate the difference in the responses to long versus short prompts, and beginner versus advanced prompts. After reviewing the model responses and statistics, we found that some of the scores were inflated due to tokenization. As a result, the responses from the models consistently contained spaces between apostrophes, such as the tokens "you're" being split into "you ' re". Because of the nature of the Dale-Chall scores formula, it gives the original response a slightly higher rating. However, responses containing many occurrences of this in a single sentence may increase the score by up to 15 percent. To fix this, we had the tokens cleaned up and re-calculated for more accurate scoring.

6.3 Hypothesis

Before testing the models, we made three hypotheses on our results:

1. The beginner model's average English readability scores will be lower than that of the advanced model's.

2. Given beginner-level prompts and advanced-level prompts, both the beginner and advanced models will generate responses with higher scores for advanced-level prompts, and lower scores for beginner-level prompts.

3. Overtime, both the beginner and advanced models will have their average response's scores converge towards the score they were split at (i.e., 6.99).

7 Evaluation

7.1 Perplexity

Perplexity is a measure of how uncertain the model is at predicting the next token in a conversation. When finetuning both the beginner and advanced models, the perplexity was calculated by taking the exponential of the evaluation loss using the PyTorch library. The perplexity was 2.96 for the beginner model, and 1.68 for the advanced model, which were ideally minimized, meaning that our models were quite certain in what words would come next in our prediction.

7.2 Statistics

Table 1 shows the statistics concerning the responses of both models given 50-word prompts. The beginner prompts resulted in average scores of 3.91 and 4.40 for the beginner and advanced models respectively. The advanced prompts resulted in responses of average scores 3.50 and 5.43 respectively. This is a 10.5 percent decrease in the average score for the beginner model when faced with advanced prompts as opposed to beginner prompts,

	Beginner model	Advanced model
Beginner prompts	3.91	4.40
Advanced prompts	3.50	5.43

Table 1: Average scores of models given 50 word prompts.

	Beginner model	Advanced model
Beginner prompts	6.61	7.46
Advanced prompts	6.71	7.10

Table 2: Average scores of models given 50 word prompts w/o anomalies.

however in the case of the advanced model there was a 23.4 percent increase in the average score when given advanced prompts compared to beginner prompts.

Due to certain responses being of a few words (number of tokens < 5), some of the scores were between 0 and 1. When taking those scores out during calculation, we are left with table 2, which contains the same data without the responses with scores between 0 and 1. The beginner prompts resulted in average scores of 6.61 and 7.46 for the beginner and advanced models respectively. The advanced prompts averaged scores of 6.71 and 7.10 for the beginner and advanced models. This now shows a 1.51 percent increase for the beginner model's average score when given advanced prompts as opposed to beginner prompts, however the advanced model had a 4.83 percent decrease in the average score when given advanced prompts as opposed to beginner prompts.

Table 3 represents the responses of both models given 200 word prompts. Due to a larger volume of 200-word prompts, these statistics should more accurately represent the model's scores. In the case of both models, their average scores increased when given advanced prompts as opposed to beginner prompts, with a 41.2 percent increase in the beginner model's average score, and a 10.1 percent increase in the advanced model's average score.

7.3 Trends and observations

This section indicates the trends observed during offline evaluation of the statistics calculated. First, we shall indicate whether or not the hypotheses hold true.

The first hypothesis holds true in all average scores calculated in the previous section, and can therefore safely be declared as true.

The second hypothesis does not hold true in both table 1 and table 2; however, in the case of table 3, which has the largest sample size, it does hold true. According to the law of large numbers, as the sample size grows, its mean gets closer to the average of the whole population. Taking into account that the sample size for table 3 is over 4 times the sample size of table 1, it is safer to take those statistics into consideration. The hypothesis does hold true for this case, and therefore can be declared as true.

The third hypothesis does not hold true for table 1 and table 3; however, in table 2, where the results that heavily affect the averages were not taken into consideration, we can see the numbers tend towards the point at which the dataset was split. It is interesting to note that in both the cases of table 1 and table 3, we can see that the advanced prompt brings the average scores closer to the middle point when compared to the beginner prompt's average scores. This hypothesis is not as easy to confirm as there is not enough conclusive evidence given the statistics, however there is not a lack of evidence for the central limit theorem. Alas, without further evidence, this hypothesis remains unconfirmed, yet not denied.

During manual evaluation, we found that 50-word prompts actually generated more cohesive responses compared to the 200-word prompts. Both models also tended to enter a cycle (i.e., start repeating words or sentences) after several turns in the conversation. However, this phenomenon was greatly reduced through the adjustment of parameters such as temperature, repetition-penalty, top-p, and top-k.

The nature of the question and answer format for the dataset used for fine-tuning resulted in 3 obvious trends when manually evaluating the model; the first was that the last sentence within the prompt generally has the most weight in response generation as it asks a question that the chatbot is supposed to respond. The second trend is that the model tends to give short responses the fine-tuning

	Beginner model	Advanced model
Beginner prompts	3.28	5.07
Advanced prompts	4.63	5.58

Table 3: Average scores of models given 200 word prompts.

dataset's responses were observed to be quite short as well. The last trend is that the model responds with declaratives more than interrogatives, possibly also due to the fact that the last sentence in our prompts is always an interrogative question.

Limitations

Since DialoGPT was trained on Reddit data, it has the possibility to produce output that could contain offensive speech. As a result, responses produced by the model could have a tendency to agree with ideas that are biased or offensive. In addition, there was limited computing power for training for more epochs and experimentation of different combinations of hyperparameters for our model. For future work, we can also choose to construct better beginner English datasets by crawling the web on EFL websites and transcripts of audio materials.

8 Conclusion

In this paper, we present an English language learning chatbot that evaluate learner's English proficiency and responds with beginner or advanced English depending on the evaluation results. We fine-tuned the DialoGPT model to train a beginner and advanced model. In testing, we found that advanced model generates more advanced responses than beginner model; moreover, given advanced prompts, both models generate more advanced responses than given beginner prompts. Furthermore, we also found that 50-word prompts elicit more coherent responses from both the beginner and advanced models than 200-word prompts.

References

Serge Bibauw, Wim Van Den Noortgate, Thomas François, and Piet Desmet. 2022. Dialogue systems for language learning: a meta-analysis. *Language Learning & Technology*, 26(1).

Michael Canale and Merrill Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1):1–47.

Richard Csaky. 2019. Deep learning based chatbot models. *arXiv preprint arXiv:1908.08835*.

Elaine K Horwitz, Michael B Horwitz, and Joann Cope. 1986. Foreign language classroom anxiety. *The Modern language journal*, 70(2):125–132.

Weijiao Huang, Khe Foon Hew, and Luke K Fryer. 2022. Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1):237–257.

Bernd Huber, Richard Davis, Allison Cotter, Emily Junkin, Mindy Yard, Stuart Shieber, Elizabeth Brestan-Knight, and Krzysztof Z. Gajos. 2019. pcit$_{d}ataset.tab$. In *Dataset for "SpecialTime: Automatically Detecting Dialogue Acts from Speech to Support Parent-Child Interaction Therapy"*. Harvard Dataverse.

Athanasios Karasimos. 2022. The battle of language learning apps: a cross-platform overview. *Research Papers in Language Teaching and Learning*, 12(1):150–166.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Nuobei Shi, Qin Zeng, and Raymond Lee. 2020. The design and implementation of language learning chatbot with xai using ontology and transfer learning. *arXiv preprint arXiv:2009.13984*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Appendix A

Example 1: User's interaction with the English Language Learning Chatbot demonstrating model switching:

User: Hi! How are you?

Beginner Bot: I'm fine, thanks.

User: Great! Do you like ice cream?

Beginner Bot: Yes, I do.

User: What is your favourite flavour?

Beginner Bot: Ice cream with chocolate sauce.

User: Awesome, mine too! Do you play League of Legends?

Beginner Bot: Yeah, I watch a lot of it.

SYSTEM MESSAGE: User's english is Advanced [8.36]

User: What is your rank on League of Legends?

Advanced Bot: I am rank 3 on the ladder.

User: Do you want to play with me tonight?

Advanced Bot: Yes.

User: I love playing Akshan, what is your favourite champion?

Advanced Bot: Kha'zhu.

B Appendix B

Example 2: User's interaction with the English Language Learning Chatbot without model switching:

User: Hello, long time no see, how is everything going?

Beginner Bot: Not bad, thanks.

User: Honestly, same. What did you do last night?

Beginner Bot: I went to the movies.

User: Cool! What did you watch?

Beginner Bot: Twilight.

User: How did you like it?

Beginner Bot: It's a very good film.

SYSTEM MESSAGE: User's english is Beginner [0.96]

User: I went to watch Spiderman last weekend, have you seen it?

Beginner Bot: Oh, yes. I saw it the first day it came out on DVD.

User: Do you like popcorn?

Beginner Bot: I like everything about it except for the cheese.

User: What is your favourite flavour?

Beginner Bot: Ice cream is my favorite.