TUTORING: Instruction-Grounded Conversational Agent for Language Learners

Hyungjoo Chae[1,3], Minjin Kim[2], Chaehyeong Kim[2],
Wonseok Jeong[3], Hyejoong Kim[3], Junmyung Lee[3], Jinyoung Yeo[1,2,3*]

[1]Department of Computer Science, Yonsei University
[2]Department of Artificial Intelligence, Yonsei University
[3]Tutoring, Market Designers Inc.
{mapoout, minjin.kim, cheris8, jinyeo}@yonsei.ac.kr, {fredric, kate, tony}@tutoring.co.kr

Abstract

In this paper, we propose TUTORING bot, a generative chatbot trained on a large scale of tutor-student conversations for English-language learning. To mimic a human tutor's behavior in language education, the tutor bot leverages diverse educational instructions and grounds to each instruction as additional input context for the tutor response generation. As a single instruction generally involves multiple dialogue turns to give the student sufficient speaking practice, the tutor bot is required to monitor and capture when the current instruction should be kept or switched to the next instruction. For that, the tutor bot is learned to not only generate responses but also infer its teaching action and progress on the current conversation simultaneously by a multi-task learning scheme. Our TUTORING bot is deployed under a non-commercial use license at https://tutoringai.com.

Introduction

With the recent success of neural dialogue generation (Shuster et al. 2022) based on pre-trained language models (Lewis et al. 2020), foreign language learning is a promising application of conversational agents in the education field. As many students experience foreign language anxiety with human tutors (called *xenoglossophobia*), the conversation with AI enables the students to easily start talking in their target language. However, prior work (Huang et al. 2017; Pham et al. 2018; Tu 2020; Shi, Zeng, and Lee 2020; Park et al. 2022) is limited to merely "chit-chat", which is not thoughtfully designed for language education. In contrast, professional and dedicated human tutors may lead the conversations through diverse and personalized educational instructions, such as guiding to read sentences for beginner level, answer questions for intermediate level, or debate on controversial issues for advanced level students.

In this demo, we present a fully data-driven concept of tutoring conversational agent, which models the tutor-student conversations without any pre-defined scenario logic and manual programming for the instructions. For that, we prepare and formulate a novel dataset/task, namely *instruction-grounded* response generation, where a sequence of educational instructions is described in natural language form and is shared between a tutor and a student for its use on conversations. Here, a straightforward implementation is to leverage the individual instructions as conditional code, each of which is concatenated to the dialogue context to make the desired tutor response and dialogue flow for language education. Despite its effectiveness, this approach is sub-optimal since the tutor agent cannot monitor how much conversation should be done for each instruction and whether the student has successfully followed the instruction or not.

To overcome this drawback, we design and leverage a set of auxiliary tasks that infer *teaching action and progress* as dialogue state, which can be jointly learned with the primary task, *i.e.*, instruction-grounded response generation. We hypothesize that such multi-task learning contributes to the injection of the auxiliary information into the generated tutor responses. Toward this goal, the action/progress labels of the auxiliary tasks are automatically annotated by in-dialogue signals such as human tutors' feedback, transition of the instructions, and the amount of dialogue turns per instruction. We empirically validate that such signals are effective to improve the response quality. To the best of our knowledge, beyond language models for chit-chat, TUTORING bot is the first instruction-based conversational agent that enables students to experience educational conversation.

System Description

Figure 1 illustrates TUTORING bot with four auxiliary tasks of inferring action and progress information in addition to the response generation task based on multi-task learning.

Tutor Response Generation with Action Codes

Let $\mathcal{X}$ and $\{\mathcal{I}_i\}_{i=1}^{N}$ be a dialogue context with T turns and a fixed sequence of N instructions, where each turn is aligned with one specific instruction $\mathcal{I}_i$. A dialogue model parameterized by θ aims to generate an appropriate response y with M tokens based on $\mathcal{X}$ and $\mathcal{I}_i$.

To allow the model to ground on instructions, we introduce action codes for the instruction and dialogue, respectively. The instruction action code y^{inst} indicates whether to move on to the next instruction $\mathcal{I}_{i+1}$, which is a special token `[Transition]` generated only when the conversation ends for the current instruction $\mathcal{I}_i$. On the other hand, the dialogue action code y^{dial} represents educational feed-

*Corresponding author.

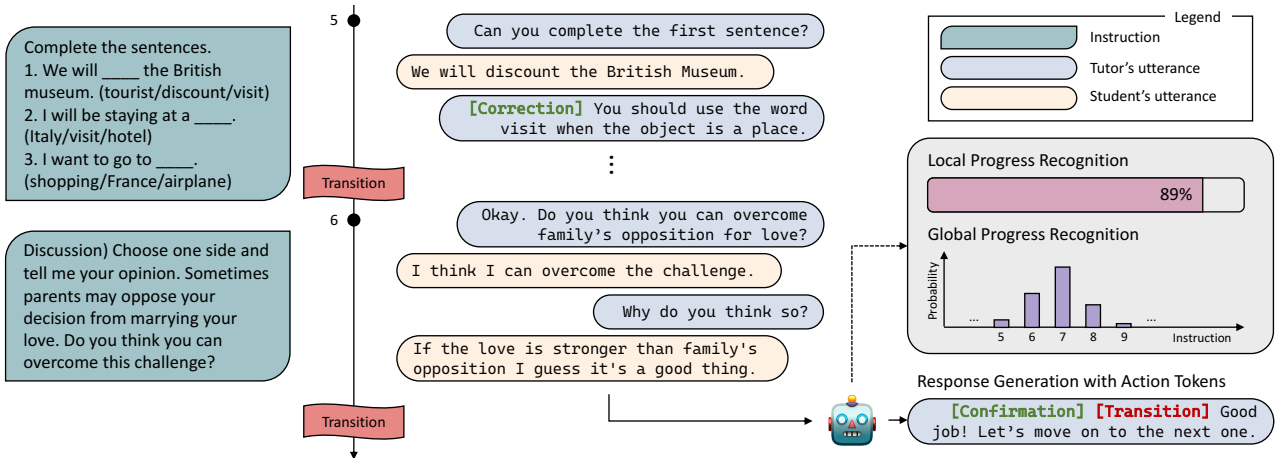

Figure 1: The system overview of TUTORING bot.

back for the given dialogue context $\mathcal{X}$, which can be either `[Correction]`, `[Confirmation]`, or `[Others]`.

Given a dialogue context $\mathcal{X}$ and its aligned instruction $\mathcal{I}_i$, the model learns to generate a response y with the action codes y^{inst} and y^{dial}. The generation loss $\mathcal{L}_{gen}$ is computed by the negative log-likelihood loss.

$$\mathcal{L}_{gen} = -\{\log p(y^{dial}|\mathcal{X},\mathcal{I}_i) + \log p(y^{inst}|\mathcal{X},\mathcal{I}_i,y^{dial}) + \sum_{j=1}^{M} \log p(y_j|\mathcal{X},\mathcal{I}_i,y^{dial},y^{inst},y_{<j})\} \quad (1)$$

Global and Local Progress Recognition

Considering the sequence of instructions from $\{\mathcal{I}_i\}_{i=1}^{N}$, we design the global progress recognition task, where the global progress y_{rec}^{glo} represents which instruction is involved in the current conversation (*i.e.*, the index i from $\mathcal{I}_i$). We further incorporate the local progress recognition task, in which the local progress y_{rec}^{loc} denotes the fraction of the number of proceeded dialogue turns over the total number of dialogue turns aligned with $\mathcal{I}_i$, thus ranging from 0 to 1.

The model learns to predict both global and local progress by the recognition loss $\mathcal{L}_{rec}$, which is defined as the sum of cross-entropy and mean squared error loss between ground truth labels y_{rec} and predicted labels $\hat{y}_{rec}$, respectively.

$$\mathcal{L}_{rec} = \mathrm{CE}(y_{rec}^{glo},\hat{y}_{rec}^{glo}) + \mathrm{MSE}(y_{rec}^{loc},\hat{y}_{rec}^{loc}) \quad (2)$$

Multi-task Learning

The dialogue model is jointly trained on the aforementioned tasks to generate instruction-grounded responses and recognize the learning progress by updating θ as follows:

$$\text{TUTORING bot}: \theta^* = \underset{\theta}{\operatorname{argmin}}\ \mathcal{L}_{gen}(\theta) + \mathcal{L}_{rec}(\theta) \quad (3)$$

Evaluation and Demonstration

Based on 11 unique instructions, we collect 1,911 dialogues of 95,343 utterances from tutor-student conversations in real world. We split the dataset with ratio 8:1:1 for the training, validation, and test sets, respectively, where the instructions are shared between the sets. We employ a pre-trained BART-large (Lewis et al. 2020) as our base model. The evaluation results in Table 1 show that incorporating both action codes and progress recognition achieves the best performance with their synergistic effect in the response generation task. Also, instruction transitions achieve 87.98% in terms of accuracy.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
RG	21.03	14.42	9.57	6.95
RG + AC	21.44	14.70	9.78	7.14
RG + PR	21.53	14.73	9.79	7.14
RG + AC + PR	**23.02**	**15.78**	**10.59**	**7.74**

Table 1: Performance of the proposed model under different configurations. RG, AC, and PR denotes response generation, action codes, and progress recognition, respectively.

We deploy TUTORING bot to online education demo by attaching Speech-to-Text[1] and Text-to-Speech[2] modules for making the system easily accessible. The expected running time of a tutoring session is about 11 minutes with 50 turns on average. Our demo video shows that TUTORING bot can properly control the real-time teaching process and timely transition within instructions with proper educational feedback. We additionally provide a debugging tool for developers to identify the generated action codes and the results of the progress recognition tasks based on visualization using Gradio (Abid et al. 2019).

As future work, we will release a public and full version of our task and dataset with 50 unique instructions and 922,446 utterances soon. Using this dataset as a testbed, the tutor bot can be advanced to a product-ready application with 1) more generalized dialogue models for diverse/unseen instructions, 2) extended auxiliary tasks for better response quality, and 3) additional feedback/correction modules on expressiveness, grammar, and pronunciation.

[1]Vosk Speech Toolkit: https://github.com/alphacep/vosk-api
[2]Coqui Tacotron2-DCA: https://github.com/coqui-ai/TTS

Acknowledgments

We would like to thank anonymous reviewers for their valuable comments. This work was partially supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022-11-0941).

References

Abid, A.; Abdalla, A.; Abid, A.; Khan, D.; Alfozan, A.; and Zou, J. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Huang, J.-X.; Lee, K.-S.; Kwon, O.-W.; and Kim, Y.-K. 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*, 151–156.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

Park, C.; Jang, Y.; Lee, S.; Park, S.; and Lim, H. 2022. FreeTalky: Don't Be Afraid! Conversations Made Easier by a Humanoid Robot using Persona-based Dialogue. In *LREC*.

Pham, X. L.; Pham, T.; Nguyen, Q. M.; Nguyen, T. H.; and Cao, T. T. H. 2018. Chatbot as an intelligent personal assistant for mobile language learning. In *Proceedings of the 2018 2nd International Conference on Education and E-Learning*, 16–21.

Shi, N.; Zeng, Q.; and Lee, R. 2020. Language Chatbot–The Design and Implementation of English Language Transfer Learning Agent Apps. In *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, 403–407. IEEE.

Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E. M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Tu, J. 2020. Learn to Speak Like A Native: AI-powered Chatbot Simulating Natural Conversation for Language Tutoring. *Journal of Physics: Conference Series*, 1693.