Towards Explainable and Satisfactory Dialogue Systems by Conversational Thought Experiments

To Eun Kim
Researcher at UCL Web Intelligence Group
`https://kimdanny.github.io`
Research Proposal

December 8, 2022

***K*eywords** Conversational AI · Information Retrieval · Task-Oriented · User Simulation · Satisfaction Estimation · Response Selection and Reranking · Explainability · Reasoning · Evaluation · Theory-of-Mind · Thought Experiments

1 Introduction

As Kim [1] has written in his dissertation, there are numerous challenges to overcome in dialogue systems - multi-turn dynamics, domain knowledge integration, data efficiency, and evaluation - and it is more so in task-oriented dialogue system (TOD), which is designed with specific domains and tasks in mind, compared to an open domain dialogue system. Due to the challenges, researchers have attempted to decompose the complex task into simpler modularised components, each of which are usually for Natual Language Understanding (NLU), Dialogue Management (DM), and Natural Language Generation (NLG). Additionally, since there can be multiple response candidates from one or more generators (running in parallel), *response selection* and *reranking* module are often added to the system.

Focused on the response selection and reranking module, I suggest using a **user simulator** in the dialogue system to have a **look-ahead signal** and building a **conversation tree** during the conversation that let us conduct **thought experiments**: prefactual, counterfactual, and hindcasting. These computational thought experiments are believed to be the key towards explainable and satisfactory conversation achieved by higher order reasoning when selecting and reranking responses. Moreover, it is important to note that the proposed research is not just limited to selection and reranking of the responses, but has a greater implication on the paradigm of action taking as well as evaluation of the system with wide-opened extensibility, such as strategic dialogue systems. Thus, Section 4 that deals with future work is equally important as other sections. From the following sections, we discuss problems that current literature has and see how the conversation tree can be used to overcome numerous challenges.

1.1 Motivation

I view conversation as a cooperative and collaborative mind game, termed as ***pragmatics*** in linguistics. I believe that conversational AI research should ground this fact as its core philosophy. Since it is a mind game, it is important for a conversational agent to generate responses not only naturally but also strategically. To achieve this, an ability to reason counterpart's mental states and realities of the surroundings, **Theory-of-Mind (ToM)**, is crucial. Computational ToM research dates back to when Turing Machine is under active research. Based on the traditional theory, we can make an assumption that there exists a Universal Turing Machine that can simulate humans' cognitive behaviours [2].

In recent years, as large language models are being constantly improved, solving a lot of downstream tasks of NLP, questions regarding whether these models have ToM have been raised. In EMNLP 2022, Sap et al. have shown that GPT-3, a state-of-the-art (SOTA) autoregressive language model (LM), significantly lacks social intelligence and ToM [3]. This made me start to think how we can give ToM ability to dialogue systems, in order for them to reason in higher order and take actions accordingly. Moreover, in that ToM can be used to figure out user's mental states during conversation, I found a connection between ToM and **user satisfaction estimation** research. If there is a **user simulator** that can predict user's future satisfaction level given dialogue history, we can leverage the hint as a *look-ahead signal* to select the best response(s) that can ensure higher satisfaction to the user. Graphical explanation and future research direction can be found in the latest talk by Kim at the NLP interest group of The Alan Turing Institute [4].

2 Related Work and Challenges

Response Selection and Reranking is a retrieval task, considering dialogue history and candidate responses as query and documents, respectively. Response reranking is important and challenging, because multiple plausible responses can be generated or selected. Inspired by the promise of pretrained language models, Henderson et al. have introduced a method which fine-tunes the pretrained model with domain-specific dialogue [5]. Based on this method, a number of models [6, 7, 8] have been setting a record in response selection benchmark datasets, such as Ubuntu Dialogue, Reddit and AmazonQA from Poly AI. However, the output responses from these models are resulted from similarity-based representation learning, not from higher order reasoning, thus lack explainability. Especially in a sophisticated system, such as TOD and strategic conversation system, e.g., CICERO [9], this can be highly problematic. Reasoning and explainability are needed not only to justify, but also to control, improve, and discover [10]. One very recent work by Hu et al. have partly alleviated the problem, focusing on reranking the overgenerated responses by two different methods: classification and similarity-based reranking [11]. However, this still misses the core philosophy of language learning, *pragmatics*, as stated in Section 1.1.

Looking Ahead. See and Manning have utilised their Alexa Prize social bot's chat history to define the taxonomy of errors that neural generative models make during the conversation and analysed when users are dissatisfied. Based on the analysis, they made a binary classification model that forecasts user's dissatisfaction given the dialogue history and the candidate response. This way, they could select the response that can avoid possible dissatisfaction of users [12]. In a similar fashion, Shin et al. have considered user's sentiment towards the generated response given the dialogue history and the candidate response. One difference is that they used this *sentiment look-ahead* as a reward under a reinforcement learning (RL) regime to provide higher reward to the response generator [13]. However, as See and Manning stated, only a minor set of user express their dissatisfaction right after the error. Therefore, only one step of sentiment look-ahead might not be sufficient. Another work from Ben-David, Carmeli, and Anaby-Tavor showed that system can improve its intent prediction task by seeing next user's utterance in advance [14]. This work is valuable in that they directly modelled user's utterance not just predicting metrics such as dissatisfaction and sentiment.

Learning from Users. System can learn from real and/or simulated users. Recent work have tried to use reals users' feedback into the retraining cycle to amend existing dataset and retrain the system [15, 16]. They both estimated users' satisfaction to learn what is desirable response. Although they could get precious user satisfaction data, the system had to be retrained to be able to improve itself. Deep Dyna-Q [17] is the first deep RL framework for TOD policy learning using user simulator. Moreover, while learning the policy, the user model is constantly updated with real users' experience. Kim and Lipani have presented a novel user simulator that 1) generates user-side utterance, 2) predicts dialogue act and 3) satisfaction level by multi-task learning [18], by using USS dataset [19]. This is the first work that included satisfaction estimation with user simulator.

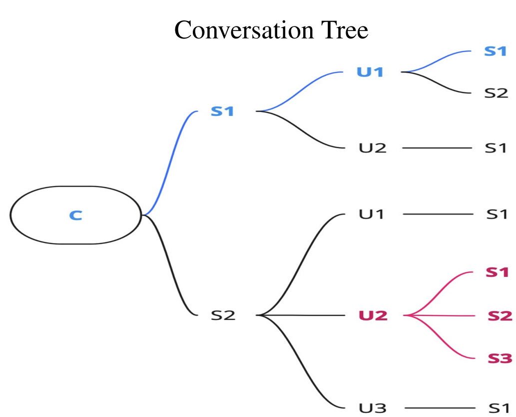

Figure 1: Conversation Tree (best viewed in colour). C, S, U denotes dialogue context, system response, and user utterance, respectively. Blue and red paths represent single path and single turn, respectively.

Evaluation of Agent and User Simulator. Most of the conversational IR systems employ offline evaluation methods, where the conversation is evaluated by the test collections, in accordance with the Cranfield paradigm. However, this method has multiple disadvantages. First, the system is limited in selecting the best response from a predefined set of candidates. Second, it can either judge a single turn or a single path of the conversation. Single turn evaluation is not ideal, because it does not consider the history of the conversation that led to the turn. Likewise, single path evaluation is not ideal, because it overlooks the other branches of the responses that could have been made by the bot during the dialogue. Figure 1 depicts the two ways of evaluation. Furthermore, these ways of evaluating the system are highly exhaustive and not scalable [20]. Evaluating the system with real users (online evaluation) is also challenging as the system has to be deployed to public. This, in turn, causes the scarcity of user satisfaction data, compared to the relative abundance of the conversation data for other training purposes. To bridge the gap, Amazon has been hosting an Alexa Prize Challenge, where participants can deploy their agents to the general public and retrieve daily feedbacks: user satisfaction score on a scale of 1 to 5 [20, 21].

Similarly, evaluation of user simulator is demanding. Although Kim and Lipani has used BLEU, ROUGE and semantic textual similarity to assess the user utterance generated from the simulator, it is still open to question how to automatically evaluate an utterance that describes new topic (a completely valid user behaviour); the scores will be extremely low according to the current automatic metrics [18]. Because of this reason, literature still relies on the human evaluation [22].

3 Response Reranking by Satisfaction Look-Ahead

As a novel way of giving reasoning ability and ensuring explainability to dialogue systems, while following the philosophy of *pragmatics*, this proposal suggests using user simulator to forecast user's future satisfaction level and use the score as a signal when reranking candidate responses. By using this *satisfaction look-ahead* signal, we expect to have higher scores in metrics of response reranking task. This 'look-ahead' approach is a prefactual thought experiment, where user simulators speculate on the outcomes (satisfaction and utterance) of each event `E` (candidate response) and the agent asks "*What will happen if* `E` *occurs?*". This aligns with our cognitive behaviour that we often think listener's perspective before formulating our speech. Additionally, for counterfactual evaluation, building a Conversation Tree that resembles Figure 1 during the conversation is proposed as a solution. Moreover, effects of depth of the future conversation tree (more than one turn of look-ahead) on reranking performance should be experimented. Although there have been a few efforts to utilise a look-ahead signal in the literature, as far as we know, there has been no attempt to use such signal for response reranking and explainability.

3.1 User Simulator for Generating User Behaviours

To predict user's satisfaction score and simulate multiple future steps of conversation, user simulator needs to be able to predict both satisfaction score and utterance (to form the conversation tree). As Kim and Lipani's user simulator [18] satisfies this requirement, we will modify/improve this model to adapt to our task.

First, current model can generate only one user utterance given the context, i.e., branching factor from the last system output is 1 in the conversation tree. This is not a problem if we intend this, but this hinders us from experimenting the stochastic nature of user behaviour. Therefore, we can let the simulator generate at most K number of user utterances by top-p sampling. This change must induce another change, where the model should predict user satisfaction in an 'after-utterance' (AU)[1] fashion, not in a 'before-utterance' (BU)[2] way, because the satisfaction of different utterance candidates are likely to be different.

3.2 Conversation Tree for Taking Satisfactory Actions.

Conversation Tree is a tree data structure consists of system responses and user utterances. It includes not only the path taken by the system but also past and future candidates. After we have the user simulator, we can let user simulator generate possible user's utterance(s) for each candidate response. This way, we formed the tree with *one-turn look-ahead*; in each node, satisfaction score is attached to the utterance. Continuously, the agent can clone the context of each branch (including the generated utterance) and branch out again with their candidate responses, followed by the same procedure of user simulation, i.e., *two-turn look-ahead*. In theory, this process of generating future conversation tree can be continued until the conversation in every branch is ended by the user simulator, i.e., N*-turn look-ahead*.

After we construct a conversation tree with a certain depth, we have to search a branch that can result in highest value, i.e. satisfaction score. Inspired by game theory and RL, we can employ Minimax or Alpha-Beta Pruning search algorithm. This tree search can also be viewed as manual temporal-difference learning, where, multiple satisfaction scores from the same parent node can be averaged to represent the expected value of an action (response) in the turn.

3.3 Experiments

To compare with the previous work [11], we can follow the same experimental set up of Hu et al. Using MultiWOZ as dataset and the best performing pretrained response selection module, we can obtain the most challenging reranking candidates. Here we can see how reranking based on the N-turn satisfaction look-ahead performs on the reranking assessment metrics (BLEU, ROUGE, METEOR). Additionally, we can experiment with varying N = 1, 2, 3, 4, ... and search algorithms to find out which N and search method normally result in the promising performance with bearable computation time. Thus, the performance should be recorded with elapsed time.

[1] AU prediction: learning a function $\mathbf{M}'(s|\mathcal{H}+u)$ after generating u from a function $\mathbf{M}(u|\mathcal{H})$, where s is satisfaction score, $\mathcal{H}$ is dialogue history ending with last system output, $+$ is concatenation, and u is user utterance.

[2] BU prediction: learning a function $\mathbf{M}(s, u|\mathcal{H})$ (same notation as AU prediction).

4 Limitations and Future Work

Need for the Cognition and Linguistics Research on Satisfaction. Our underlying assumption is that user satisfaction is the ultimate goal of dialogue systems so that optimisation of it can guarantee a satisfactory conversation. In this perspective, estimating user's satisfaction can free us from considering other metrics, such as naturalness and task success rate. However, the hypothesis has never been rigorously tested, and it is not easy to assess user's satisfaction, as the concept of satisfaction is complex and can be varied depending on the type of conversation. Therefore, from cognition and linguistics point of view, we have to define 1) what a 'good' conversational flow is, and 2) what factors lead to the satisfactory conversation in different environment. This research can be done by letting human subjects converse on the same topic but with a number of chat bots that have their own conversational strategy. For each experiment, we need to capture when and how the subjects express their satisfaction or dissatisfaction. Only after formalising the concept of satisfaction, we can better estimate user satisfaction.

Need for More Advanced User Simulator. A limitation of the proposed research is that the reranking performance heavily depends on the performance of user simulator. Therefore, more sophisticated simulator is crucial. One way to improve the simulator is modularisation and personalisation. As the conversational agent was implemented as a set of modules, specialised in their own task, simulator can be made in the same way. Also, user simulators can largely benefit from user adaptation. It is because each individual has their own personal preference, persona, stochastic behaviour, and level of ability to learn the system and form expectations [20]. For example, What is deemed to be a commonsense may not be commonsensical to certain people. Bearing this in mind, we can build a personal knowledge graph (PKG) [23] that interacts with each module of the simulator. In the NLU component, we can refine the outcome of the original NLU model based on the PKG. This can simulate the situation where user cannot understand what conversational assistant for professional domain is saying due to difficult terminologies. The same logic applies to other component, such as NLG. Balog explains further on how to incorporate *User Model*, *Interaction Model*, and *Mental Model* into the user simulator [20]. Although, a user simulator with this level of complexity may not be needed nor easy/fast to compute when interacting with the conversational agent for look-ahead computation, the direction of the research is still valid as we need much finer user simulation to achieve our goal.

Conversational Thought Experiments. Other than the prefactual reasoning, we can evaluate the agent by counterfactual evaluation, i.e., asking "What if event E' had happened instead of event E?". Evaluating such a way can drive the agent to be more robust to various scenarios of conversation.
One of the challenges mentioned was evaluation of user simulator. To test the performance of the simulator, we can use hindcasting, i.e., running the user simulator after the chat between real user had happened. This way, we can check whether the generated user utterances are similar to the utterances of real user given the same context.

References

[1] To Eun Kim. "Multi-Task Neural User Simulator for Task Oriented Dialogue System". In: (2022). URL: `https://drive.google.com/file/d/1tA9DimDM9Mz5cWdbelZgsIP8D63evW2n/view`.

[2] Michael Rescorla. *The computational theory of mind*. Feb. 2020. URL: `https://plato.stanford.edu/entries/computational-mind/`.

[3] Maarten Sap et al. *Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs*. 2022. DOI: `10.48550/ARXIV.2210.13312`. URL: `https://arxiv.org/abs/2210.13312`.

[4] To Eun Kim. "Can Dialogue Systems Have Theory-of-Mind Ability?" In: (2022). URL: `https://drive.google.com/file/d/1gpAGb6hy3Drn7JIq4P1FfD-5ZrDg4m8m/view?usp=share_link`.

[5] Matthew Henderson et al. "Training Neural Response Selection for Task-Oriented Dialogue Systems". In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2019. DOI: `10.18653/v1/P19-1536`. URL: `https://aclanthology.org/P19-1536`.

[6] Samuel Humeau et al. "Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring". In: *International Conference on Learning Representations*. 2020. URL: `https://openreview.net/forum?id=SkxgnnNFvH`.

[7] Matthew Henderson et al. "ConveRT: Efficient and Accurate Conversational Representations from Transformers". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Nov. 2020. DOI: `10.18653/v1/2020.findings-emnlp.196`. URL: `https://aclanthology.org/2020.findings-emnlp.196`.

[8] Janghoon Han et al. "Fine-grained Post-training for Improving Retrieval-based Dialogue Systems". In: *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, June 2021. DOI: `10.18653/v1/2021.naacl-main.122`. URL: `https://aclanthology.org/2021.naacl-main.122`.

[9] Anton Bakhtin et al. “Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning”. In: *Science* (2022). DOI: 10.1126/science.ade9097. eprint: https://www.science.org/doi/pdf/10.1126/science.ade9097. URL: https://www.science.org/doi/abs/10.1126/science.ade9097.

[10] Amina Adadi and Mohammed Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018). DOI: 10.1109/ACCESS.2018.2870052.

[11] Songbo Hu et al. *Reranking Overgenerated Responses for End-to-End Task-Oriented Dialogue Systems*. 2022. DOI: 10.48550/ARXIV.2211.03648. URL: https://arxiv.org/abs/2211.03648.

[12] Abigail See and Christopher Manning. “Understanding and predicting user dissatisfaction in a neural generative chatbot”. In: *Proc. of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, July 2021. URL: https://aclanthology.org/2021.sigdial-1.1.

[13] Jamin Shin et al. “Generating empathetic responses by looking ahead the user’s sentiment”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020.

[14] Eyal Ben-David, Boaz Carmeli, and Ateret Anaby-Tavor. *Improved Goal Oriented Dialogue via Utterance Generation and Look Ahead*. 2021. arXiv: 2110.12412 [cs.CL].

[15] Braden Hancock et al. “Learning from Dialogue after Deployment: Feed Yourself, Chatbot!” In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2019. DOI: 10.18653/v1/P19-1358. URL: https://aclanthology.org/P19-1358.

[16] Weiyan Shi et al. *When Life Gives You Lemons, Make Cherryade: Converting Feedback from Bad Responses into Good Labels*. 2022. DOI: 10.48550/ARXIV.2210.15893. URL: https://arxiv.org/abs/2210.15893.

[17] Baolin Peng et al. “Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning”. In: *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, July 2018. DOI: 10.18653/v1/P18-1203. URL: https://aclanthology.org/P18-1203.

[18] To Eun Kim and Aldo Lipani. “A Multi-Task Based Neural Model to Simulate Users in Goal Oriented Dialogue Systems”. In: *Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’22. Madrid, Spain: Association for Computing Machinery, 2022. DOI: 10.1145/3477495.3531814. URL: https://doi.org/10.1145/3477495.3531814.

[19] Weiwei Sun et al. “Simulating User Satisfaction for the Evaluation of Task-oriented Dialogue Systems”. In: *Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (July 2021). DOI: 10.1145/3404835.3463241. URL: http://dx.doi.org/10.1145/3404835.3463241.

[20] Krisztian Balog. “Conversational AI from an Information Retrieval Perspective: Remaining Challenges and a Case for User Simulation”. In: *Proc. of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021*. Ed. by Omar Alonso et al. Vol. 2950. CEUR Workshop Proceedings. CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2950/paper-03.pdf.

[21] Jerome Ramos et al. “Condita: A state machine like architecture for multi-modal task bots”. In: *Alexa Prize TaskBot Challenge Proceedings*. 2022. URL: https://www.amazon.science/alexa-prize/proceedings/condita-a-state-machine-like-architecture-for-multi-modal-task-bots.

[22] Weiyan Shi et al. “How to Build User Simulators to Train RL-based Dialog Systems”. In: *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Nov. 2019. DOI: 10.18653/v1/D19-1206. URL: https://aclanthology.org/D19-1206.

[23] Krisztian Balog and Tom Kenter. “Personal Knowledge Graphs: A Research Agenda”. In: *Proc. of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR ’19. Association for Computing Machinery, 2019. DOI: 10.1145/3341981.3344241. URL: https://doi.org/10.1145/3341981.3344241.