Target-Guided Open-Domain Conversation Planning

Yosuke Kishinami[1]   **Reina Akama**[1,2]   **Shiki Sato**[1]   **Ryoko Tokuhisa**[1]
Jun Suzuki[1,2]   **Kentaro Inui**[1,2]
[1]Tohoku University   [2]RIKEN
yosuke.kishinami.q8@dc.tohoku.ac.jp
{akama,shiki.sato.d1,tokuhisa,jun.suzuki,inui}@tohoku.ac.jp

Abstract

Prior studies addressing target-oriented conversational tasks lack a crucial notion that has been intensively studied in the context of goal-oriented artificial intelligence agents, namely, *planning*. In this study, we propose the task of Target-Guided Open-Domain Conversation Planning (TGCP) task to evaluate whether neural conversational agents have goal-oriented conversation planning abilities. Using the TGCP task, we investigate the conversation planning abilities of existing retrieval models and recent strong generative models. The experimental results reveal the challenges facing current technology.

1 Introduction

Neural conversational agents have achieved great successes in recent years, and various methods have been proposed to generate informative responses, e.g., the use of knowledge (Zhao et al., 2020; Wu et al., 2020), personality (Li et al., 2016; Zhang et al., 2018), emotional considerations (Rashkin et al., 2019; Zhong et al., 2020), and large-scale models (Zhang et al., 2020; Adiwardana et al., 2020; Roller et al., 2021; Thoppilan et al., 2022). One hot topic in this research area is to develop proactive behavior in agents. For example, Tang et al. (2019) proposed the task of Target-Guided Open-Domain Conversation, in which an agent is required to actively lead a conversation to a predefined target word. Wu et al. (2019) proposed a task that uses a knowledge graph to actively lead a conversation to a target entity. Several studies have implemented these target-oriented task settings (Dai et al., 2019; Qin et al., 2020; Yuan and An, 2020; Zhong et al., 2021; Zhu et al., 2021). However, these prior studies all lack *planning*, a crucial notion that has been intensively studied in the context of goal-oriented artificial intelligence (AI) agents (Norvig and Russell, 1995; Kuijpers and Dockx, 1998; Stent et al., 2004; Walker et al.,
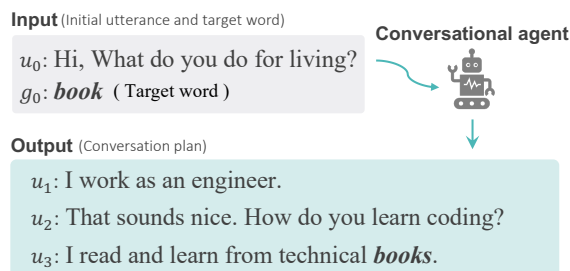

Figure 1: Overview of the TGCP task.

2007, etc.) and has also been introduced in neural conversational agents (Botea et al., 2019; Jiang et al., 2019a,b). In other words, these studies do not explicitly consider the generation of a multiple-step plan to achieve a target.

Given this background, in this study, we propose the **Target-Guided Open-Domain Conversation Planning** (henceforth, **TGCP**) task such that an agent's planning ability in goal-oriented conversations can be assessed. The TGCP task is to produce a plan that leads a conversation to a given target, as illustrated in Figure 1. The point is to consider the task of producing a conversation plan for several utterances ahead, which we first address in the aforementioned context of Target-Guided Open-Domain Conversation. Furthermore, we also propose modeling the planning process by simulating the user's succeeding utterances using the model of the agent itself; namely, the agent converses with itself (i.e., self-conversation) to search for potential conversation paths that achieve the goal. This task setting is not the same as a real-world setting, in which an agent is required to plan a conversation while uncertain of the user's future utterances. However, planning in the self-conversation setting can be considered a prerequisite capability for a planning-aware goal-oriented conversational agent. TGCP works as a framework to evaluate an agents' prerequisite ability for conversation planning without employing human subjects; this can abstract

660

away the hard-to-control human factors from experiments (e.g., some human subjects may not be as cooperative as others).

This paper has three major contributions: (1) We propose the TGCP task as a framework to assess the prerequisite ability of a model for goal-oriented conversation planning. (2) We conduct a set of experiments on the TGCP framework using several existing retrieval-based neural models and recently proposed strong generative neural models of conversational agents. (3) Our experimental results reveal the challenges facing current technology. The evaluation codes and the test set used in the experiments are available.[1]

2 Target-Guided Open-Domain Conversation Planning

We introduce the task of Target-Guided Open-Domain Conversation Planning, the TGCP task for short, that is to evaluate whether the agents have goal-oriented conversation planning abilities. In this section, we describe the task definition and the evaluation metrics.

2.1 Task definition

Figure 1 shows an overview of the TGCP task. We define the goal given to the agents as a word (e.g., *dog*, *pizza*, *coffee*). Given a target word g_0 and an initial utterance u_0, TGCP requires agents to make an entire conversation plan $(u_1, \ldots, u_N)$, whose last utterance u_N, which consists of M words, contains the target word g_0, namely, $u_N = (w_{N,1}, \ldots, w_{N,M})$, and $w_{N,m} = g_0$ for any $m \in M$. This task has the same input/output format as the human-agent conversation task proposed by Tang et al. (2019). However, these task setups differ in terms of whether or not a human conversational partner is involved. In TGCP, agents generate for all utterances in the entire conversation.

2.2 Evaluation metrics

The evaluation is performed based on three objectives: whether the target word is mentioned (**achievement ratio**), whether the utterance transitions in the conversation are natural (**transition smoothness**), and how likely the conversation is to actually occur (**conversation probability**). We believe that satisfying these three perspectives is important in goal-oriented conversation planning.

For example, given a target word *computer* and an initial utterance "What sports do you like?," the utterance like *I love computer.* achieves the target, but it is not natural as a conversation, and such an interaction rarely occurs. Likewise, the utterance like "*I don't like sports because my friend who likes sports broke my computer.*" is a natural transition and achieves the target, but it is likely to rarely occur in an actual conversation. We believe that an agent's generation of such utterances does not indicate the agent's planning ability. We can automatically calculate the achievement ratio based on whether the target word itself is mentioned or not.[2] We also consider transition smoothness and conversation probability to be manually evaluated.[3]

3 Experiments

Using the proposed TGCP, we investigate the conversation planning ability of several major existing dialogue models and recent deep neural network (DNN) based dialogue models.

3.1 TGCP settings

Dataset. As a dataset for the TGCP task, we prepared $1,000$ pairs consisting of an initial utterance and a target word, i.e., (u_0, g_0). We created these pairs by randomly extracting from a set of the first utterances and a set of keywords extracted from subsequent utterances in the ConvAI2 dataset.[4]

Evaluation. As described in Section 2.2, in TGCP, the conversation plans generated by models are evaluated by target achievement ratio, transition smoothness, and conversation probability. The target achievement ratio was calculated automatically. To avoid infinite conversations that never reached the target, we set the maximum number of turns to 8.[5] For transition smoothness and conversation probability, we evaluated them manually using Amazon Mechanical Turk.[6] For each model, randomly sampled 100 conversation plans were rated by native English speakers. We eliminated

---

[2]Tang et al. (2019) considers mentioning synonyms as the task achievement; however, Zhong et al. (2021) points out that synonyms are unreliable to measure the task achievement. Implementation details are provided in Appendix A.1.

[3]Empirical analyses on the relationship between these two metrics are provided in Appendix A.2.

[4]This follows existing analogous work (Tang et al., 2019; Qin et al., 2020; Zhong et al., 2021). In addition, we removed the keywords not covered by ConceptNet.

[5]The same setting as previous studies (Tang et al., 2019; Qin et al., 2020; Zhong et al., 2021).
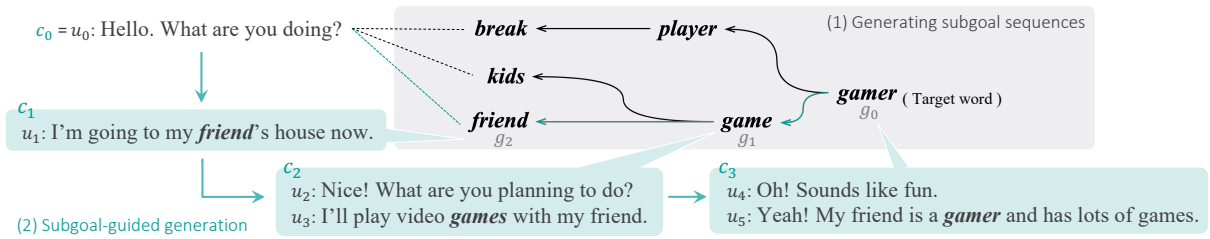
[6]https://www.mturk.com/

Figure 2: Subgoal-guided conversation plan generation with BLENER+PREDES..

low-quality workers using attention checks. Five workers rated each conversation on a five-point Likert scale for transition smoothness (5 is *Strongly good* and 1 is *Strongly bad*) and conversation probability (5 is *Frequently* and 1 is *Rarely*).[7]

3.2 Existing models

We prepared the following seven existing dialogue models employed in Target-Guided Open-Domain Conversation: Wu et al. (2017)'s RETRIEVAL, Tang et al. (2019)'s RETRIEVAL-ST., PMI, NEURAL, and KERNEL, Qin et al. (2020)'s DKRN, and (Zhong et al., 2021)'s CKC.[8] All models except RETRIEVAL are retrieval dialogue models that infer the keyword to mention immediately after each turn of the conversation on the fly and then determine the next response based on it and the conversational history. RETRIEVAL is a retrieval dialogue model that does not infer the keyword but determine the next response only based on conversational history.

3.3 Recent DNN-based models

In addition, we prepared the latest generative model that combines the DNN-based powerful dialogue model, BLENDER (Roller et al., 2021), and a novel strategy for pre-designing keyword sequences to given g_0 (PREDES.). Our BLENDER+PREDES. (Figure 2) is a newly designed model. Note that this model is new because the task is new, and that there should be many ways to design models for TGCP. Still, we believe that testing the performance of a specific model such as BLENDER+PREDES. on TGCP can help investigate the nature of TGCP. In our BLENDER+PREDES., we first generated the keyword sequences, hereinafter it called *subgoal sequence*, using ConceptNet5 (Speer et al., 2017). Specifically, we acquired the series of n concepts that are passed when tracing the edges of the knowledge graph from the concept representing the target word g_0 to the concept related

to the initial utterance u_0 as a subgoal sequence $G = g_0, g_1, \cdots, g_{n-1}$. n is the length of the subgoal sequence including the target word g_0. This allows preventing cases that cannot get closer to the target than a certain point because of selecting a locally optimal solution. After generating the subgoal sequence, we generated a sequence of partial conversations $C = c_0, c_1, \cdots, c_n$ using BLENDER as follows:

$$c_i = f(g_{n-i}, (c_0, \cdots, c_{i-1})) \quad (1 \le i \le n) \quad (1)$$

Where, c_i denotes a partial conversation that follows the previous partial conversation c_{i-1} and ends up with the utterance where the subgoal g_{n-i} appears. $f(\cdot)$ is a function that returns a partial conversation to the given previous conversations and a subgoal.

PREDES. settings. We set $n = 3$, i.e., we generated subgoal sequences by tracing ConceptNet up to three levels.[9,10] Among the subgoal sequences generated from ConceptNet, we retained the 30 subgoal sequences in which the end of the subgoal sequence g_{n-1} was the most related to the given first utterance u_0. We calculated the relatedness as the cosine similarity between the SIF embedding (Arora et al., 2017) of u_0 and GloVe word vector (Pennington et al., 2014) of g_{n-1}. Among the conversation plans generated from the 30 subgoal sequences, we selected the conversation plan with the highest average probability of generating partial conversations by BLENDER as the final output.

Training of BLENDER. We used the Blenderbot 3B implemented by huggingface transformer

---

[7]Concrete instructions are provided in Appendix C.
[8]Implementations details are given in Appendix B.1.

[9]We excluded all stopwords in the NLTK and spaCy libraries to comprehensively exclude unnecessary words. In addition, we excluded the concepts for which the score calculated by wordfreq (Speer et al., 2018) was lower than the score of the target word.
[10]We empirically confirmed the validity of this setting (Appendix B.3).

Model	Subgoal	Conversation	Achievement	#Turns	Smoothness	Probability
RETRIEVAL (Wu et al., 2017)	-	retrieval	0.034	3.71	3.52	3.37
RETRIEVAL-ST. (Tang et al., 2019)	on-the-fly	retrieval	0.851	5.04	3.50	3.29
PMI (Tang et al., 2019)	on-the-fly	retrieval	0.531	4.97	3.33	3.17
NEURAL (Tang et al., 2019)	on-the-fly	retrieval	0.535	2.83	3.14	3.00
KERNEL (Tang et al., 2019)	on-the-fly	retrieval	0.596	2.79	3.24	3.05
DKRN (Qin et al., 2020)	on-the-fly	retrieval	**0.968**	2.91	3.28	3.12
CKC (Zhong et al., 2021)	on-the-fly	retrieval	0.353	3.60	2.81	2.69
BLENDER (Roller et al., 2021)	-	generative	0.024	5.04	3.99	**3.90**
BLENDER+CKC	on-the-fly	generative	0.247	7.00	3.90	3.71
BLENDER+PREDES.	pre-design	generative	0.425	6.29	**4.05**	**3.90**
Human	-	-	1.000	3.50	4.11	3.89

Table 1: Performance of dialogue models on the TGCP task.

as BLENDER.[11] Because BLENDER is a model that generates a response based on the conversational history, we fine-tuned it to use as f which generates a partial conversation based on previous partial conversations and a subgoal. we used the ConvAI2 processed by Zhong et al. (2021) as a training data for BLENDER. We prepared the training data by randomly splitting a single conversation into an input and an output consisting of multiple utterances and then concatenating a word extracted randomly from the output-side utterances (i.e., keywords) to the input utterances.[12] We finally obtained 117,877 pairs as training set and 6,425 pairs as validation set. The hyperparameters are provided in Appendix B.2.

3.4 Ablation models

To analyze the effectiveness of the pre-design strategy, we also prepared Blender without any conversational strategy (BLENDER), and with an on-the-fly strategy using existing models. Specifically, we employed the strategy of CKC as the comparison on-the-fly strategy, which is known to be the highest performance method in TGC (Zhong et al., 2021) (BLENDER+CKC). For both models, Blender is the same as BLENDER+PREDES.. But note that blender, without any conversation strategy, does not concatenate keywords with inputs for training and inference.

3.5 Results

Table 1 shows the evaluation results on TGCP. To provide the human upper bound performance, we also had three workers perform TGCP on 50 pairs

randomly selected from the dataset described in Section 3.1 (Human).

Achievement ratio. The achievement ratios of the retrieval models tended to be high. In particular, the achievement ratio of DKRN was comparable to that of humans. The generative models had lower achievement ratios. However, BLENDER+PREDES. improved the achievement ratio compared with BLENDER+CKC, whose subgoal strategy is the same as that of CKC. This result means that replacing the on-the-fly subgoal strategy with the pre-design strategy is effective in improving the target achievement ratios of generative models.

Smoothness & probability. The retrieval models have lower values of transition smoothness and conversation probability than humans. Table 2 shows a conversation plan example generated by DKRN, whose achievement ratio was the highest of all the compared models. In the example, the transition between u_1 and u_2 is clearly unnatural, although the model achieved to mention the target word.[13] The transition smoothness and conversation probability of the generative models were higher than those of the retrieval models. In particular, BLENDER+CKC significantly outperformed CKC in these metrics. Therefore, using powerful DNN-based generation models improves the transition smoothness and conversation probability of the conversation plans. Table 3 shows a conversation plan example generated by BLENDER+PREDES., whose transition smoothness was the highest of all the compared models. In this example, BLENDER+PREDES. generated a natural conversation along an appropriately gener-

---

[11] https://github.com/huggingface/transformers
[12] We extracted the keywords by following Zhong et al. (2021). The pairs that failed to extract keywords from the output utterances were excluded from the training data.

[13] An additional example is provided in Appendix D.2.

u_0	hey how is it going ?
u_1	i'm doing ok . i have mass this week (*school*: 0.64)
u_2	i just got done sewing a new shirt (*shirt*: 1.00)

Table 2: Part of the conversation plan by an existing model (DKRN). The elements in parentheses are keywords predicted by the model and the similarity score between the keyword and the target. (Target: **shirt**)

ated subgoal sequence.

Overall. The TGCP task revealed the planning abilities of well-known retrieval models and newly prepared generative models. The retrieval models tended to have high achievement ratios but low transition smoothness and conversation probability, while the opposite was true for the generative models. These results show the trade-off between achievement ratio and the naturalness of conversation plans that current technology is facing. On the other hand, the generative model with a pre-design subgoal strategy (BLENDER+PREDES.) improved the achievement ratio compared with the generative model with an on-the-fly strategy (BLENDER+CKC) ensuring its high transition smoothness and conversation probability. This implies that improving the achievement ratios of generative models by refining their subgoal strategies is an effective approach to overcome the trade-off.

3.6 Discussion: Number of conversation turns

We found that generative models behave critically differently from humans regarding the number of turns to reach targets, while their transition smoothness and conversation probability were comparable to those of humans. The average numbers of turns to reach targets of BLENDER+CKC and BLENDER+PREDES. were much larger than that of humans.[14] This result indicates that humans efficiently achieved TGCP with fewer turns while ensuring high conversation probability.[15] Making an agent has a strong conversation planning ability like a human can be one of our challenges in the future.

4 Conclusion

We have proposed the TGCP task as a platform for assessing the conversation planning ability of

u_0	hello how are you ?
u_1	not too bad, how about you?
u_2	i am good thanks for asking
u_3	what do you do for a living, if you don't mind me asking? i am a nurse
u_4	that's cool. i work in a grocery store.
u_5	do you like it
u_6	yeah, it pays the bills, but i want to be a dental hygienist
u_7	wow that's a great career choice. how long have you been doing that
u_8	for as long as i can *remember*
u_9	what do you like to do in your spare time
u_{10}	i love to take *pictures* and photography is a hobby of mine
u_{11}	what kind of pictures do you take?
u_{12}	mostly *landscapes*, i love nature

Table 3: Conversation plan by BLENDER+PREDES. (Target: **landscape**). The predicted subgoal sequence is *remember* → *picture* → **landscape**.

a dialogue model. Through this task setting, we have presented a first study for assessing the present neural conversational models' abilities for multiple-utterance planning, abstracting away the hard-to-control potential human factors. While the reported experiments cover only the task of Target-Guided Open-Domain Conversation (Tang et al., 2019), the idea of TGCP is expected to be applicable to a wider range of goal-oriented conversation tasks. Using TGCP, we revealed that the dialogue models with current technology have difficulty planning conversations to achieve given goals while ensuring the naturalness of the conversation. The experimental results also showed that refining the subgoal strategies for generative models might be an effective method to overcome this trade-off. We plan to research methods to solve this task setting with higher performance.

Acknowledgments

---

[14]We counted the number of turns of the conversations where the target words are mentioned.

[15]A generated example is provided in Appendix D.1.

Ethical considerations

This paper honors the ACL Code of Ethics. This study uses existing dataset, preprocessed versions of ConvAI dataset Tang et al. (2019); Zhong et al. (2021), which we believe does not involve any ethical concerns. We used these data as training data for dialogue models and keyword prediction models and to randomly extract subsets for use as inputs for our generation task; we do not believe this involves any ethical concerns. This study includes manual work; For human evaluation, we hired crowdworkers and paid appropriate rewards for their labor (corresponding to a $14.40 for an hourly wage).

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. In *aiXiv preprint arXiv:2001.09977*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Adi Botea, Christian Muise, Shubham Agarwal, Oznur Alkan, Ondrej Bajgar, Elizabeth Daly, Akihiro Kishimoto, Luis Lastras, Radu Marinescu, Josef Ondrej, Pablo Pedemonte, and Miroslav Vodolan. 2019. Generating Dialogue Agents via Automated Planning. In *The Second AAAI Workshop On Reasoning And Learning For Human-Machine Dialogues (DEEP-DIAL)*.

Zelin Dai, Weitang Liu, and Guanhua Zhan. 2019. Multiple Generative Models Ensemble for Knowledge-Driven Proactive Human-Computer Dialogue Agent. In *aiXiv preprint arXiv:1907.03590*.

Zhuoxuan Jiang, Jie Ma, Jingyi Lu, Guangyuan Yu, Yipeng Yu, and Shaochun Li. 2019a. A General Planning-Based Framework for Goal-Driven Conversation Assistant. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 9857–9858.

Zhuoxuan Jiang, Xian Ling Mao, Ziming Huang, Jie Ma, and Shaochun Li. 2019b. Towards End-to-End Learning for Efficient Dialogue Agent by Modeling Looking-ahead Ability. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 133–142.

Bart Kuijpers and Kris Dockx. 1998. An Intelligent Man-Machine Dialogue System Based on AI Planning. *Applied Intelligence*, 8(3):235–245.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003.

Peter Norvig and Stuart J Russell. 1995. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jinghui Qin, Zheng Ye, Jianheng Tang, and Xiaodan Liang. 2020. Dynamic Knowledge Routing Network for Target-Guided Open-Domain Conversation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 8657–8664.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y. Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5370–5381.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y. Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 4444–4451.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable Sentence Planning for Complex Information Presentations in Spoken Dialog Systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 79–86.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-Guided Open-Domain Conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5624–5634.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee Huaixiu, Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang,

Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel, Morris Tulsee, Doshi Renelito, Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le Google. 2022. LaMDA: Language Models for Dialog Applications. In *aiXiv preprint arXiv:2201.08239*.

Marilyn Walker, Amanda Stent, Francois Mairesse, and Rashmi Prasad. 2007. Individual and Domain Adaptation in Sentence Planning for Dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5811–5820.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive Human-Machine Conversation with Explicit Conversation Goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3794–3804.

Yu Wu, Wei Wu, Chen Xing, Zhoujun Li, and Ming Zhou. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–505.

Hao Yuan and Jinqi An. 2020. Multi-Hop Memory Network with Graph Neural Networks Encoding for Proactive Dialogue. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence (ICCAI)*, pages 24–29.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 270–278.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-Guided Neural Conversational Model. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 14568–14576.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards Persona-Based Empathetic Conversational Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Yutao Zhu, Jian Yun Nie, Kun Zhou, Pan Du, Hao Jiang, and Zhicheng Dou. 2021. Proactive Retrieval-based Chatbots based on Relevant Knowledge and Goals. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2000–2004.

A Details of the TGCP Task

A.1 Calculation of Target Achievement Ratio

The achievement judgment was based on whether the target word itself was mentioned (Zhong et al., 2021); however, we found that there were several cases in which the achievement was judged to be a failure even though the target word was appropriately mentioned. Therefore, we modified the script[16] for judging the achievement such that these cases would be judged as achievements.

A.2 Relationship between Transition Smoothness and Conversation Probability

We investigated the correlation between transition smoothness and conversation probability of the experiment in Section 3, and found that Pearson's correlation coefficient was 0.828, which indicates a high correlation. Therefore, it appears that conversation probability is contained within transition smoothness, at least in our experiment. Therefore, evaluating transition smoothness may indicate the approximate tendency of conversation probability.

B Model Implementations and Setups

B.1 Existing Models

We used publicly available codes by their authors to implement the existing models.[17,18] To train the response selection models and the keyword prediction models, we used the same dataset and setups as described in their papers: CKC used the ConvAI2 dataset processed by Zhong et al. (2021), and the other models used the ConvAI2 dataset processed by Tang et al. (2019).

B.2 Training parameters of BLENDER.

To train BLENDER, we set the batch size to 32, the learning rate to 7.0×10^{-6}, the warmup steps to 100, the evaluation steps to 1,000, and the number of updates to 50,000. The other parameters were set to the default configuration of the huggingface transformer. We used the model at the validation loss minimum point for conversation planning.

B.3 Length of subgoal sequence.

We empirically confirmed the validity of tracing ConceptNet up to the three levels using the fol-

u_0	Not a big fan of talking face to face . How about you?
u_1	Me too. I prefer texting.
u_2	I truly understand. People love to comment on your behaviors when talking face to face. But talking online does not have such problems.
u_3	It sounds like you have experienced such comments. What did people accuse you of?
u_4	Well, I am a ***vegetarian***, but they said vegetarians are incomprehensible. Rude people, aren't they?

Table 4: Conversation plan generated by a human. (Target: ***vegetarian***)

lowing procedure. First, we qualitatively checked the subgoal sequences generated by PREDES. and found that the subgoal sequences with relatedness scores of approximately 0.6, indicating a connection with the initial utterance, were naturally connected with the initial utterance. Then, we investigated how much ConceptNet need to be searched to generate subgoal sequences with a score of approximately 0.6. As a result, we confirmed that by tracing ConceptNet up to three steps, the average score of the subgoal sequences of the search results was 0.653, which exceeded 0.6. Therefore, we conclude that the three-step search is reasonable.

C Instructions for Human Evaluation

Figure 3 shows the instructions given to Amazon Mechanical Turk workers concerning the evaluation of transition smoothness and conversation probability.

D Generated Conversation Plans

D.1 Human-generated Conversation Plan

Table 4 shows a conversation plan generated by a human. We confirmed that human could plan natural conversations that achieved their target despite the short number of turns.

D.2 Dead-ended Case

Table 5 shows an example where RETRIEVAL-ST. did not achieve its target. After the keyword *catch* is selected in utterance u_6, RETRIEVAL-ST. continued to generate the same utterance repeatedly from u_7 to the end (u_{16}). Ultimately, RETRIEVAL-ST. could not reach the target.

---

[16]https://github.com/zhongpeixiang/CKC/blob/master/util/data.py
[17]https://github.com/James-Yip/TGODC-DKRN
[18]https://github.com/zhongpeixiang/CKC

Evaluation of the transition smoothness and probability of the conversation

You are provided conversations between person (A) and person (B).
Please answer the following two questions.
- How smooth are the transitions during the conversation?
- How likely is the conversation to occur?

This HIT contains 11 evaluations (#01-#11).

Task description

First, you are given the conversation between person A and person B

> A: I like cookies. What about you?
> B: I like it. I like cookies that taste like vegetables.
> A: Really? I don't like vegetables, but I like games about growing them.
> B: I see. I didn't know there was such a game.

Then, you are asked the following two question:

Question1: How smooth are the transitions during the conversation?
Please rate transition smoothness during the conversation on the five-point scale:
- **Score 5: Strongly Good**
- **Score 4: Good**
- **Score 3: Unsure**
- **Score 2: Bad**
- **Score 1: Strongly Bad**

In the above example, Overall the utterance transitions are smooth, so the score is 5.

Question2: How likely is the conversation to occur?
Please rate conversation probability on the five-point scale:
- **Score 5: Frequently**
- **Score 4: Often**
- **Score 3: Sometimes**
- **Score 2: Occasionally**
- **Score 1: Rarely**

In the above example, The score is 1 because conversations about "*cookies that taste like vegetables*" and "*games about growing vegetables*" are rarely likely to occur.

* Please read and evaluate the meaning of the entire utterance, without considering minor issues such as grammar or spelling errors.

Figure 3: Evaluation instructions given to Amazon Mechanical Turk workers.

u_0	Today I am annoyed about not being able to drive.
u_1	That's *horrible*. My friend is also paralyzed. (*horrible*: 0.20)
u_2	I am sorry to *hear* of your injury. (*hear*: 0.43)
u_3	I feel like dump and skeleton but just *wasting* my hours here. (*waste*: 0.48)
u_4	I don't understand *people* with egos. (*people*: 0.50)
u_5	I think happiness is an illusion which *money* can help with. (*money*: 0.53)
u_6	I don lie, its bad when you get *caught*. (*catch*: 0.54)
u_7	I guess that's why the constantly keep calling me. (*catch*: 0.54)
u_8	I guess that's why the constantly keep calling me. (*catch*: 0.54)
$\vdots$	
u_{16}	I guess that's why the constantly keep calling me. (*catch*: 0.54)

Table 5: Part of the conversation plan by an existing model (RETRIEVAL-ST.). The elements in parentheses are keywords predicted by the model and similarity score between the keyword and target. (Target: ***chase***)