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., 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
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.

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. We also consider transition smoothness and conversation probability to be manually evaluated.

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., \(\langle u_0, g_0 \rangle\). We created these pairs by randomly extracting from a set of the first \(5,000\) utterances and a set of keywords extracted from subsequent utterances in the ConvAI2 dataset.

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. For transition smoothness and conversation probability, we evaluated them manually using Amazon Mechanical Turk. For each model, randomly sampled 100 conversation plans were rated by native English speakers. We eliminated

---

1The evaluation codes and the test set are available at https://github.com/y-kishinami/TGCP

2Tang 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.

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

4This 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.

5The same setting as previous studies (Tang et al., 2019; Qin et al., 2020; Zhong et al., 2021).
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 \(G = g_0, g_1, \ldots, g_{n-1}\). \(n\) 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 \(C = c_0, c_1, \ldots, c_n\) using BLENDER as follows:

\[
c_i = f(g_{n-i}, (c_0, \ldots, c_{i-1})) \quad (1 \leq i \leq n)
\]

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.

---

Figure 2: Subgoal-guided conversation plan generation with BLENDER+PREDES..
as Blender. 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. 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).

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

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

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. 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-
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)

| u0   | hey how is it going?
| u1   | i'm doing ok. i have mass this week (school: 0.64)
| u2   | i just got done sewing a new shirt (shirt: 1.00)

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

| u0   | hello how are you?
| u1   | not too bad, how about you?
| u2   | i am good thanks for asking
| u3   | what do you do for a living, if you don't mind me asking? i am a nurse
| u4   | that's cool. i work in a grocery store.
| u5   | do you like it
| u6   | yeah, it pays the bills, but i want to be a dental hygienist
| u7   | wow that's a great career choice. how long have you been doing that
| u8   | for as long as i can remember
| u9   | what do you like to do in your spare time
| u10  | i love to take pictures and photography is a hobby of mine
| u11  | what kind of pictures do you take?
| u12  | mostly landscapes, i love nature

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 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

We would like to thank all anonymous reviewers for their invaluable comments. This work was partly supported by JSPS KAKENHI Grant Numbers JP21J22383, JP22K17943, JST Moonshot R&D Grant Number JPMJMS2011.
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


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


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\textsuperscript{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.\textsuperscript{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 \(7.0 \times 10^{-6}\), 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 following 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 \textit{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.

\textbf{Table 4: Conversation plan generated by a human. (Target: vegetarian)}

\begin{tabular}{l}
\hline
\(u_0\) & 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 \textit{vegetarian}, but they said vegetarians are incomprehensible. Rude people, aren’t they? \\
\hline
\end{tabular}
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 questions:

Question 1: 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.

Question 2: 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.

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Text</th>
<th>Score</th>
<th>Keywords</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>u₁₀</td>
<td>Today I am annoyed about not being able to drive.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₁</td>
<td>That’s horrible. My friend is also paralyzed. (horrible: 0.20)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₂</td>
<td>I am sorry to hear of your injury. (hear: 0.43)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₃</td>
<td>I feel like dump and skeleton but just wasting my hours here. (waste: 0.48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₄</td>
<td>I don’t understand people with egos. (people: 0.50)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₅</td>
<td>I think happiness is an illusion which money can help with. (money: 0.53)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₆</td>
<td>I don lie, its bad when you get caught. (catch: 0.54)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₇</td>
<td>I guess that’s why the constantly keep calling me. (catch: 0.54)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u₈</td>
<td>I guess that’s why the constantly keep calling me. (catch: 0.54)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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)