Multi-Turn Target-Guided Topic Prediction with Monte Carlo Tree Search

Jingxuan Yang, Si Li* and Jun Guo
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{yjx, lisi, guojun}@bupt.edu.cn

Abstract

This paper concerns the problem of topic prediction in target-guided conversation, which requires the system to proactively and naturally guide the topic thread of the conversation, ending up with achieving a designated target subject. Existing studies usually resolve the task with a sequence of single-turn topic prediction. Greedy decision is made at each turn since it is impossible to explore the topics in future turns under the single-turn topic prediction mechanism. As a result, these methods often suffer from generating sub-optimal topic threads. In this paper, we formulate the target-guided conversation as a problem of *multi-turn topic prediction* and model it under the framework of Markov decision process (MDP). To alleviate the problem of generating sub-optimal topic thread, Monte Carlo tree search (MCTS) is employed to improve the topic prediction by conducting long-term planning. At online topic prediction, given a target and a start utterance, our proposed MM-TP (MCTS-enhanced MDP for Topic Prediction) firstly performs MCTS to enhance the policy for predicting the topic for each turn. Then, two retrieval models are respectively used to generate the responses of the agent and the user. Quantitative evaluation and qualitative study showed that MM-TP significantly improved the state-of-the-art baselines.

1 Introduction

Although impressive efforts have been made to integrate background knowledge into the conversation systems (Fang et al., 2018; Qin et al., 2019; Liu et al., 2018), existing open-domain conversation systems still suffer from creating generic response (Yi et al., 2019) and struggle to perform engaging conversations (Ram et al., 2018). Moreover, there exists a strong demand in real-world applications to integrate the goals and strategies into the open-domain conversation systems, to make them achieve some specific goals such as recommending an item or accomplishing nursing goals. Faced with these problems, target-guided open-domain conversation (Tang et al., 2019; Sevegnani et al., 2021) has attracted increasing research attentions.

Topic	Response	*Target : job*
Turn: 1	A: Hi, how are you today? U: Pretty good. Chilling and listening to some music.	
Turn: 2 *listen*	A: Cool, I'm *listening* to some old rock music. U: Awesome, which kind of music is your favorite?	
Turn: 3 *music*	A: Hank Williams. I grew up on country *music*. U: Cool. Not much of a singer.	
Turn: 4 *job*	A: When I was a child, I had to get a *job* to help my parents.	

Figure 1: An example conversation conducted between the single-turn topic prediction conversational system agent (A) and user (U).

Different from traditional open-domain conversation, target-guided open-domain conversation requires the system to proactively and naturally guide the conversational thread, and end up with recommending a target item or mentioning a target word. Existing studies (Tang et al., 2019; Qin et al., 2020; Zhong et al., 2020) usually resolve the task with a sequential of single-turn topic predictions and response generations. At each turn, the model firstly selects a topic from the candidate topic set based on the history context, and then retrieves response according to the selected topic. Since the single-turn topic prediction mechanism has no ability to plan the topics in the future turns, greedy decision has to be made at each turn. As a result, these methods usually suffer from generating sub-optimal topic threads.

Figure 1 illustrates an example conversation between user and the Kernel agent (Tang et al., 2019), which utilizes single-turn topic prediction model to select topics. At the third turn, the sub-optimal topic "music" was selected. Though it is strongly

* Corresponding author

Proceedings of the 18th International Conference on Natural Language Processing, pages 324–334
Silchar, India. December 16 - 19, 2021.

relevant to the topic "listen" in the second turn, it is irrelevant to the final target topic "job". The example verified that the greedy decisions in the single-turn topic prediction cannot naturally guided the conversation to achieve the target.

To deal with the issue, we propose to formulate target-guided conversation as a multi-turn topic prediction problem, and model it with Markov decision process (MDP). In the MDP, the environment is responsible for collecting the conversational history as the states, and the conversational system agent is responsible for selecting action as topic for each turn. Inspired by the reinforcement learning method of AlphaGo Zero (Silver et al., 2017), we utilize Monte Carlo tree search (MCTS) to make a long-term planning by considering the topics in the future turns and then generate topic for the current turn. Given a pre-defined target topic and a randomly selected start utterance, the proposed model, referred to as MM-TP (MCTS-enhanced MDP for Topic Prediction), iteratively generates topic sequence and guides the conversation to achieve the target topic. At each turn, MCTS is firstly utilized to enhance the raw policy and predict the topic of this turn. Two retrieval models are then respectively employed to generate the responses of the agent and the user. In this way, the problem of generating sub-optimal topic threads could be alleviated by the MCTS at a certain extent.

We conducted experiments on two popular target-guided open-domain conversation benchmarks. Quantitative results show that MM-TP outperformed the state-of-the-art baselines by achieving the target more accurately and providing more smooth topic transition. Qualitative study also show that our MM-TP improved baseline methods by making long-term planning of the topics. The major contributions of the paper are three-fold:

- To the best of our knowledge, it is the first time that the target-guided conversation is formalized as a multi-turn topic prediction problem and solved under the framework of MDP.
- We adapt the traditional MCTS for the target-guided open-domain conversation, to alleviate the sub-optimal topic threads generation problem by performing long-term planning.
- The proposed MM-TP model outperformed the baseline methods in terms of achieving the targets more accurately and making more smoothly topic transition.

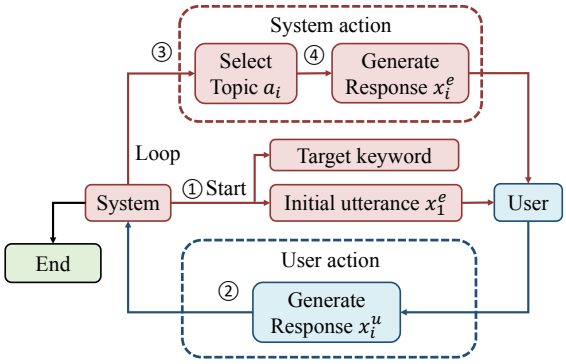

Figure 2: Workflow of Multi-turn Target-guided Open-domain Conversation

2 Task Definition: Multi-turn Target-guided Topic Prediction

As shown in Figure 2, a multi-turn target-guided open-domain conversation system starts with randomly selecting a specific target topic and the start utterance (step 1) by the simulator. The user generates an appropriate response (step 2). Then, the system repeats several conversational turns before achieving the ends. At each turn, the system first accesses to conversational history utterances and predicts a topic (step 3) satisfying both transition smoothness and target achievement. Then the agent and user generate responses respectively according to the predicted topic (step 4 and step 2). During the conversation, the target word is only presented to the agent and is unknown to user. The system consists of two components which are topic prediction module and response generation module.

Formally, let's use $\mathcal{A}$ and $\mathcal{X}$ to denote the sets of candidate target topics and responses, respectively. Following the practices in (Tang et al., 2019; Qin et al., 2020), each target topic $a \in \mathcal{A}$ is defined as a word/phrase (i.e., an entity name or a common noun), and the candidate utterance set $\mathcal{X}$ is derived from the PersonaChat corpus (Zhang et al., 2018). Suppose that the agent e starts a conversation (1st turn) with utterance $\boldsymbol{x}_1^e$ and its target topic is a^*. The user retrieval model $\mathcal{G}^u$ generates a response $\boldsymbol{x}_1^u$. Then, at each turn $i \in \{2, \cdots, m\}$, the topic prediction module takes previous utterance context $\mathrm{X}_i = \{\boldsymbol{x}_1^e, \boldsymbol{x}_1^u, \cdots, \boldsymbol{x}_{i-1}^u\}$ as input and outputs the predicted topic a_i. Then, the retrieval model $\mathcal{G}^u$ for user u and $\mathcal{G}^e$ for system agent e select a response from the candidate set $\mathcal{X}$ respectively. As an appropriate measurement of the success rate, the target is regarded as achieved when the predicted topic a_m is similar enough to the target topic a^*.

3 The Proposed Model: MM-TP

3.1 Model overview

In this work, we focus on formulating Multi-turn Target-guided Topic Prediction as an MDP and utilizing the MCTS-enhanced policy to select the topic for each turn with a long-term planning. For the response generation process, we utilize the simulator constructed in (Tang et al., 2019), and employ kernel-based retrieval model as $\mathcal{G}^e$ and conventional retrieval model as $\mathcal{G}^u$ to generate responses by agent and user respectively. Figure 3 illustrates the architecture of the proposed MCTS-enhanced MDP for Topic Prediction (MM-TP) model. Given the target word a^* and the start utterance $\boldsymbol{x}_1^e$, our model iterates several turns for guiding the conversation thread. For each turn, MM-TP first applies MCTS to select topic for the current turn, and then utilizes the retrieval models $\mathcal{G}^e$ and $\mathcal{G}^u$ to generate agent and user response respectively.

3.2 MDP formulation of Multi-turn Target-guided Topic Prediction

MM-TP models the Multi-turn Target-guided Topic Prediction as a process of sequential decision making with MDP, in which each time step corresponds to a conversational turn. The states, actions, transition function, rewards, value function and policy function of the MDP are defined as:

States $\mathcal{S}$: The state of each turn is defined as a tuple $s_t = [\mathrm{X}_t = \{\boldsymbol{x}_1^e, \boldsymbol{x}_1^u, \cdots, \boldsymbol{x}_{t-1}^u\}, \mathrm{Y}_t = \{a_1, \ldots, a_{t-1}\}]$ where X_t is the sequence of contextual utterances and Y_t is the sequence of predicted topics in previous $t-1$ turns. For the second turn, the state is initialized as $s_2 = [\{\boldsymbol{x}_1^e, \boldsymbol{x}_1^u\}, \emptyset]$, where $\{\boldsymbol{x}_1^e, \boldsymbol{x}_1^u\}$ denotes the randomly selected start utterance and the first response of user. $\emptyset$ denotes the empty topic sequence.

Actions $\mathcal{A}$: At each turn t, the $\mathcal{A}(s_t) \subseteq \mathcal{Y}$ is the set of actions the agent can choose from, which means the action $a_t \in \mathcal{A}(s_t)$ is the predicted topic $a_t \in \mathcal{Y}$ for the current turn.

Transition function T: The transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is defined as: $s_{t+1} = T(s_t, a_t) = T([\mathrm{X}_t, \mathrm{Y}_t], a_t) = [\mathrm{X}_{t+1}, \mathrm{Y}_t \oplus a_t]$, where $\oplus$ appends the selected action a_t to Y_t. At each turn t, based on state s_t, the system predicts a topic a_t for this turn, moves to the turn $t+1$ and transits the state to the next state s_{t+1}: first, the conversational utterance context X_t is updated by appending the generated agent and user responses; second, the system adds the predicted topic to the end of Y_t, outputting a new topic sequence.

Rewards $\mathcal{R}$: The reward is defined to reflect: (1) target achievement $\mathcal{R}_{ta}$: we calculate the similarity between the predicted topic of each turn and the target to determine whether the topic has achieved the target; (2) local smoothness $\mathcal{R}_{ls}$: we calculate the average WordNet similarity between topics of adjacent turns to measure the topic transition smooth; (3) target similarity $\mathcal{R}_{ts}$: we calculate the similarity difference between the adjacent topics and the target, to make the predicted topic in each turn is more similar to that in the preceding turns.The overall reward is defined as the weighted summation these three parts as:

$$\mathcal{R} = \alpha \cdot \mathcal{R}_{ta} + \beta \cdot \mathcal{R}_{ls} + \gamma \cdot \mathcal{R}_{ts}$$

where α, β, γ are weight parameters for three kinds of rewards respectively.

Value function V: The value function V is a scalar evaluation which is learned to estimate the the quality of topic assignments and fit the real evaluation measure. In this work, we utilize a hierarchical GRU network to map the context X_t to a real vector, and then define the value function as a nonlinear transformation of the weighted sum of the MLP's outputs $g(s)$ and the current candidate action in one-hot representation a_t as:

$$\mathrm{V}(s) = \sigma(\langle \mathbf{W}_v g(s), a_t \rangle),$$

where $\mathbf{W}_v \in \mathbb{R}^{|\mathcal{A}(s)| \times |g(s)|}$ is the weight vector to be learned during training. $\langle \cdot, \cdot \rangle$ is dot product operation, and $\sigma(\cdot)$ is the nonlinear sigmoid function. The context state $g(s)$ is obtained as:

$$g(s) = \mathrm{MLP}(l(s)),$$

$$l(s) = [\mathrm{HierarchalGRU}(\mathrm{X}_t)].$$

The hierarchical GRU network takes in a sequence of contextual utterances $\mathrm{X}_t = \{\boldsymbol{x}_1^a, \boldsymbol{x}_1^u, \cdots, \boldsymbol{x}_{t-1}^u\}$ and utilizes the word-level GRU to encode each utterance and output a representation of the utterance. Then, the sequence of utterance representations are fed into a utterance-level GRU for obtaining a conversational context representation $l(s)$.

Policy function $\mathbf{p}$: The policy function $\mathbf{p}(s)$ takes the context representation $g(s)$ as input and outputs a distribution over all possible actions $a \in \mathcal{A}(s)$, in which each element represents the probability of selecting this keyword as:

$$p(a|s) = \mathrm{softmax}(\mathrm{U}_p g(s)),$$

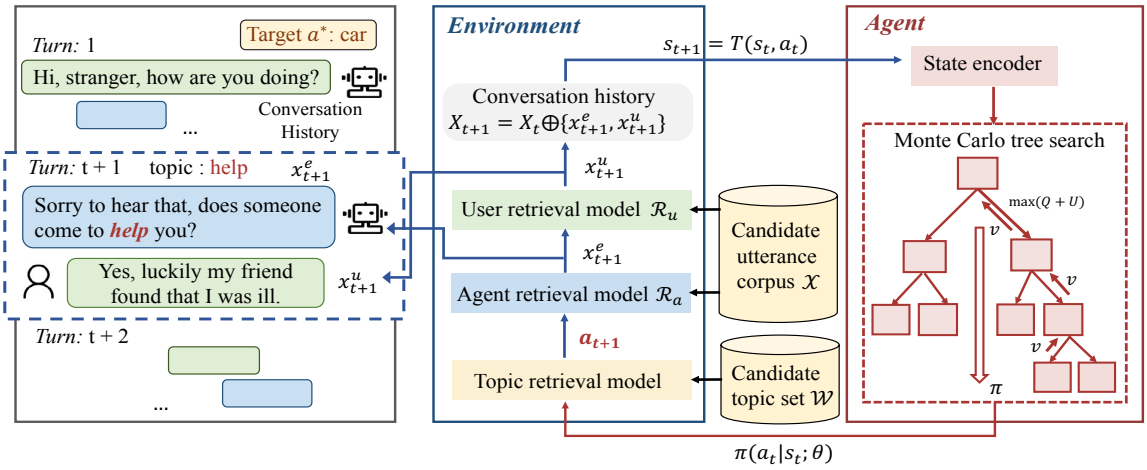

Figure 3: Overview of MM-TP. The agent guides the conversation to achieve the target by multi-turn topic prediction, which is formulated as an MDP process. For each turn, the model first encodes the previous conversation context, and then updates the search policy π to predict the topic for the next turn.

where $\mathrm{U}_p \in \mathbb{R}^{|\mathcal{A}(s)|\times|g(s)|}$ is the parameter. The policy function is obtained as:

$$\mathbf{p}(s) = \langle p(a_1|s), \cdots, p(a_{|\mathcal{A}(s)|}|s) \rangle . \quad (1)$$

At online topic sequence prediction stage, the environment collects the conversational history utterances and the predicted topic sequence as the states s_t, and then pass them to the system agent. Once received states, the agent firstly encodes them through the hierarchical GRUs and then performs MCTS to update the search policy $\boldsymbol{\pi}$, guided by the policy function $\mathbf{p}$ and value function V. The updated policy $\boldsymbol{\pi}$ is used to select action as the predicted topic for this turn.

3.3 Improve raw policy with MCTS

Predicting the topic for each turn with the raw policy $\mathbf{p}$ (Eq. 1) only considers the past states and often leads to sub-optimal results. To alleviate the issue, we conduct lookahead search with MCTS for each turn and output a improved search policy $\boldsymbol{\pi}$ to select the topic.

Specifically, MCTS takes a root node s_{R}, value function V and policy function $\mathbf{p}$ as input, and iterates K times to output a improved search policy $\boldsymbol{\pi}$ which selects a topic for the current turn. Each tree node corresponds to an MDP state. Each edge $e(s, a)$ stores an action value $\mathrm{Q}(s, a)$, visit count $\mathrm{N}(s, a)$ and prior probability $\mathrm{P}(s, a)$. For each iteration, the raw policy $\mathbf{p}$ is improved by four steps: (1) Selection: Each iteration starts from the root node s_{R} and iteratively selects a topic for each turn to maximize action value plus a bonus as $a_t = \arg\max_a(\mathrm{Q}(s_t, a) + \lambda \mathrm{U}(s_t, a))$, where $\lambda \geq 0$ is the tradeoff coefficient, and the bonus $\mathrm{U}(s_t, a) = p(a|s_t)\frac{\sqrt{\sum_{a' \in \mathcal{A}(s_t)} \mathrm{N}(s_t, a')}}{1+\mathrm{N}(s_t, a)}$ is proportional to the prior probability but decays with repeated visits to encourage exploration. (2) Evaluation and expansion: When the traversal reaches a leaf node s_{L}, the node is evaluated with the value function V. Then, the leaf node s_{L} is expanded by constructing edge from it to the node $\mathrm{T}(s_{\mathrm{L}}, a)$, corresponding to each action $a \in \mathcal{A}(s_t)$. (3) Back-propagation and update: At the end of evaluation, the action values and visit counts of all traversed edges are updated, while the prior probability $\mathrm{P}(s, a)$ is kept unchanged. (4) Calculate the improved search policy: After iterating K times, the improved search policy $\pi(a|s_{\mathrm{R}})$ corresponds to each $a \in \mathcal{A}(s_{\mathrm{R}})$ for the current root node s_{R} is calculated based on the visit counts $\mathrm{N}(s_{\mathrm{R}}, a)$ of the edges starting from s_{R}. The details of MCTS process is described in Algorithm 1.

3.4 Model training and inference

MM-TP has some parameters $\boldsymbol{\Theta}$ to learn including $\mathbf{W}_v, \mathbf{W}_g, \mathrm{U}_p, b_g$ and parameters in hierarchical GRUs. Suppose we are given N target topics and ground-truth topic threads that achieved the corresponding target topics: $\mathcal{D} = \{(a^{*(n)}, \mathrm{Y}^{(n)})\}_{n=1}^{N}$. Firstly, the parameters $\boldsymbol{\Theta}$ of the model are initialized to random weights in $[-1, 1]$. Then for each sample $(a^*, \mathrm{Y}) \in \mathcal{D}$, a topic sequence is predicted as: for each turn, the MCTS is executed and a topic a_t is selected by the search policy $\boldsymbol{\pi}_t$. The topic prediction process terminates after m turns,

Algorithm 1 TreeSearch

Input: root s_{R}, Value function V, policy function $\mathbf{p}$, search times K

for $k = 0$ to $\mathrm{K} - 1$ **do**
  $s_{\mathrm{L}} \leftarrow s_{\mathrm{R}}$
  {Selection}
  while s_{L} is not a leaf node **do**
    $a \leftarrow \arg\max_a(\mathrm{Q}(s_t, a) + \lambda \mathrm{U}(s_t, a))$
    $s_{\mathrm{L}} \leftarrow$ child node pointed by (s_{L}, a)
  end while
  {Evaluation and expansion}
  $v \leftarrow \mathrm{V}(s_{\mathrm{L}})$ {simulate v with V}
  for all $a \in \mathcal{A}(s_{\mathrm{L}})$ **do**
    Expand e to $s = [s_{\mathrm{L}}.\mathrm{X}_{t+1}, \mathrm{Y}_t \oplus \{a\}]$
    $e.\mathrm{P} \leftarrow p(a|s_{\mathrm{L}}); e.\mathrm{Q} \leftarrow 0; e.\mathrm{N} \leftarrow 0$
  end for
  {Back-propagation}
  while $s_{\mathrm{L}} \neq s_{\mathrm{R}}$ **do**
    $s \leftarrow$ parent of s_{L}
    $e \leftarrow$ edge from s to s_{L}
    $e.\mathrm{Q} \leftarrow \frac{e.\mathrm{Q} \times e.\mathrm{N} + v}{e.\mathrm{N}+1}$
    $e.\mathrm{N} \leftarrow e.\mathrm{N} + 1; s_{\mathrm{L}} \leftarrow s$
  end while
end for
for all $a \in \mathcal{A}(s_{\mathrm{R}})$ **do**
  $\pi(a|s_{\mathrm{R}}) \leftarrow \frac{e(s_{\mathrm{R}},a).\mathrm{N}}{\sum_{a' \in \mathcal{A}(s_{\mathrm{R}})} e(s_{\mathrm{R}},a').\mathrm{N}}$
end for
return $\boldsymbol{\pi}$

and a topic sequence $\hat{\mathrm{Y}} = \{a_1, \ldots, a_m\}$ is outputted. The overall evaluation metric r of $\hat{\mathrm{Y}}$ is calculated according to the success rate of the target achievement. The data generated at each turn $\mathrm{E} = \{(s_t, \pi_t)\}_{t=1}^{m}$ and the reward $\mathcal{R}$ are utilized as the signal for adjusting the value function. The training objective is to minimize the error between the predicted value $\mathrm{V}_{(s_t)}$ and evaluation metric r, and to maximize the similarity between the raw policy $\mathbf{p}_{(s_t)}$ and the search policy $\boldsymbol{\pi}_t$ as:

$$l(\mathrm{E}, r) = \sum_{t=1}^{|\mathrm{E}|}((\mathrm{V}_{(s_t)} - r)^2 + \sum_{a \in \mathcal{A}(s_t)} \pi_t(a|s_t) \log \frac{1}{p(a|s_t)}). \quad (2)$$

Algorithm 2 shows the details of the training process. The inference process of the MM-TP model is similar to the training stage. Given the selected target topic, the state is initialized as $s_2 = [\mathrm{X}_1, \mathrm{Y}_1]$. For each turn $t \in \{2, \cdots, \mathrm{m}\}$, the agent receives

Algorithm 2 Train MM-TP model

Input: Labeled data D, learning rate η, search time K, pre-defined number of turn m

Initialize Θ as random values in $[-1, 1]$
repeat
  for all $(\mathrm{X}, \mathrm{Y}) \in \mathrm{D}$ **do**
    $s_2 = [\mathrm{X}_1, \mathrm{Y}_1]; \mathrm{E} \leftarrow \emptyset$
    for $t = 1$ to m **do**
      $\pi \leftarrow$ TreeSearch $(s, \mathrm{V}, \boldsymbol{\pi}, \mathrm{K})$
      $a = \arg\max_{a \in \mathcal{A}_{(s)}} \pi(a|s)$
      $\mathrm{E} \leftarrow \mathrm{E} \oplus \{(s, \pi)\}$
      $s \leftarrow [s.\mathrm{X}_{t+1}, s.\mathrm{Y}_t \oplus \{a\}]$
    end for
    $r \leftarrow$ Metric$(\mathrm{Y}, s.\mathrm{Y}_m)$
    $\Theta \leftarrow \Theta - \eta \frac{\partial \ell(\mathrm{E}, r)}{\partial \Theta}$ {see ℓ in Eq. 2}
  end for
until converge
return Θ

the state $s_t = [\mathrm{X}_t, \mathrm{Y}_t]$ and updates the search policy $\boldsymbol{\pi}$ with MCTS. Then, MM-TP selects an action a_t for this turn and moves to the next turn whose state becomes $s_{t+1} = [\mathrm{X}_{t+1}, \mathrm{Y}_{t+1}]$.

3.5 Implementation details

We adapt the MCTS algorithm according to our task. Following existing practice (Tang et al., 2019; Qin et al., 2020), in order to guide the topic thread to achieve the target keyword, we shrink the action space in each conversational turn. Specifically, we mask the candidate topics which have been selected in preceding turns, and the candidates that are not as similar to the target as the topics in preceding turns. The tree nodes corresponding to these masked nodes thus will not be achieved during the update process of search policy $\boldsymbol{\pi}$. Moreover, we also load the parameters of pre-trained single-turn topic prediction model (Tang et al., 2019) to initialize the policy and value network, and the parameters are also updated during training process.

4 Experiments

4.1 Experimental settings

Datasets: We evaluated the performance of MM-TP on two popular conversation benchmarks: Target-Guided PersonaChat dataset (TGPC) and Chinese Weibo Conversation dataset (CWC). The TGPC dataset (Tang et al., 2019) is derived from the PersonaChat corpus which covers a abroad range of topics. Following (Tang et al., 2019),

Dataset	CWC		TGPC	
	Train	Test	Train	Test
#Conversations	824,742	45,763	8,939	500
#Utterances	1,104,438	60,893	101,935	5,317
#Keyword types	1,760	1,760	2,678	1,571

Table 1: Statistics of training and test sets on two conversation benchmarks.

Model	TGPC		CWC	
	Succ.(%)	Turns	Succ.(%)	Turns
Retrieval	7.16	4.17	0	-
Retrieval-Stgy	47.80	6.7	44.6	7.42
PMI	35.36	6.38	47.4	5.29
Neural	54.76	4.73	47.6	5.16
Kernel	62.56	4.65	53.2	4.08
DKRN	89.0	5.02	84.4	4.20
MM-TP	**91.23**	**4.82**	**86.3**	**4.15**

Table 2: Results of our MM-TP and baseline conversation systems in terms of successful rate ("Succ.%") and average turns of target achievement ("Turns").

we take 500 conversations with relatively frequent keywords as the test set. The CWC dataset (Qin et al., 2020) is a Chinese conversational dataset that derived from corpus crawled from Sina Weibo paltform. It matches the real-word scenarios better and more efficient for the model to learn dynamic topic transition. The statistics of these two benchmarks are reported in Table 1.

Baselines: Existing target-guided open-domain conversation systems are used as baselines: (1) Retrieval (Wu et al., 2017) is a conventional retrieval-based chitchat system that used to provide reference performance in terms of different metrics; (2) Retrieval-Stgy (Tang et al., 2019) which augments the above Retrieval system with the target-guided strategy and permits the system to retrieve a response containing more than one keyword; (3) PMI (Tang et al., 2019) which constructs a keyword pairwise matrix, and calculates the association between keywords by pointwise mutual information; (4) Neural (Tang et al., 2019) which utilizes a neural network to encode the conversation history and then employs a prediction layer to select a keyword for the next turn. (5) Kernel (Tang et al., 2019) which firstly measures the similarity between the current keyword and candidate keywords, and then utilizes a kernel layer to predict the candidate probability distribution; (6) DKRN (Qin et al., 2020) which uses the semantic knowledge relations among candidate keywords to mask the candidates uncorrelated to the conversational history.

Training Details: Following (Tang et al., 2019; Qin et al., 2020), we used GloVe (Pennington et al., 2014) to initialize word embeddings for English conversation corpus TGPC and Baidu Encyclopedia Word2Vec (Li et al., 2018) to initialize word embeddings for Chinese conversation corpus CWC. The number of conversational turns m was set as 8. The hierarchical GRU network utilized a hidden layer of 200 units. We used the AdaGrad (Duchi et al., 2011) optimizer to update the parameters during the training process, with a learning rate η as 0.001. The search time K in MCTS was set to 1600, and the tradeoff coefficient λ was set to 80.0. Two retrieval systems $\mathcal{G}_e$ and $\mathcal{G}_u$ were implemented with the toolkit Texar (Hu et al., 2019).

4.2 Self-play simulation evaluation

We first conducted simulation-based evaluation of our MM-TP and baseline systems in the multi-turn target-guided conversation setup. Same as (Tang et al., 2019; Qin et al., 2020), we employed the conventional retrieval system to play the role of human. The baseline models and our MM-TP played the role of system agent aiming to guide the conversation to achieve the target topic. During the training process, we generated the ground-truth topic threads by iteratively appending the keyword sequences from the consecutive single-turn keywords prediction samples in existing work (Tang et al., 2019). In the testing phrase, the simulator randomly selected a target from the candidate topic set and the start utterance from the corpus. The experiment was evaluated by measuring the success rate of achieving the target (**Succ.%**), and the average number of turns used to reach the target (**Turns**). The target topic is considered as achieved when any item of the predicted topic sequence takes a similarity score with the target higher then 0.9, measured by WordNet (Fellbaum and Miller, 1998).

Table 2 reports the results of our MM-TP as well as the baselines on TGPC and CWC. From the results, we can see that our proposed model outperformed the baselines in terms of success rate on both of the datasets. We attribute this to that MM-TP takes a long-term planning to select the topic by considering the topics in next several turns. Moreover, the average turns of MM-TP to achieve the target is comparable to baseline methods since

Model	Smoothness
Retrieval-Stgy	0.08
PMI	0.21
Neural	0.25
Kernel	0.23
DKRN	0.31
MM-TP	**0.35**

Table 3: Results of MM-TP and baseline methods in terms of transition smoothness.

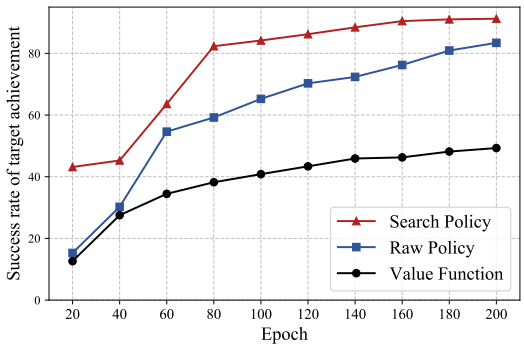

Figure 4: Success rate w.r.t. training process, when the actions were selected with raw policy $\mathbf{p}$, search policy $\boldsymbol{\pi}$, and value function V.

the long-term planning explores an optimized topic thread to achieve the target.

4.3 Effects of Monte Carlo tree search

The search policy $\boldsymbol{\pi}$ usually performs better than the raw policy $\mathbf{p}$ since MCTS is employed to consider the topics in next several turns. Except the policies, the value function V can also be used to select topic at each turn. To explore the effectiveness of these three components, we applied them to predict the topic sequence on the test set respectively after every 20 training epochs during the online training phrase, and records the average success rate of target achievement. Figure 4 illustrates the success rate curves of the raw policy $\mathbf{p}$, search policy $\boldsymbol{\pi}$, and value function V. We can see that: (1) The topic sequences generated by the search policy $\boldsymbol{\pi}$ achieves higher success rate of target achievement than that generated by the raw policy $\mathbf{p}$, which demonstrates that MCTS improved the raw policy. (2) The results predicted by both $\boldsymbol{\pi}$ and $\mathbf{p}$ are better than results predicted by the value function V. The reason is that the raw policy $\mathbf{p}$ and value network V greedily select topic at each conversational turn, which makes the results are not as good as that predicted by the search policy $\boldsymbol{\pi}$. Moreover, the quality of topic assignment is not easy to estimate by value function.

4.4 Transition smoothness evaluation

We further explore how our MM-TP accomplishes transition smoothness, which is also an important objective of target-guided conversation for measuring how naturally the conversation is guided. We evaluate our proposed model and baseline methods in terms of transition smoothness. Specifically, the transition smoothness (**Smoothness**) of each model is calculated by the average WordNet information content similarity between topics in adjacent turns. Table 3 shows the results of transition smoothness of our proposed MM-TP and baseline methods. We can see that MM-TP achieves higher transition smoothness compared with baseline systems. We contribute this to that the baseline methods are only constrained to select a topic at each turn which is strictly more closer to the target topic than those in preceding turn, while the transition smoothness between the topics in adjacent turns is overlooked. The proposed MM-TP improves these methods by modeling the transition smoothness between topics in adjacent turns as local rewards, and the performance of transition smoothness can be controlled by adjusting the weight parameter β of local smoothness.

4.5 Qualitative study

To dive a bit deeper and look at the performance of our MM-TP on topic sequence prediction, we compare the examples outputted by different conversational systems, and the results are shown in Figure 5. The three agents are given the same target topic and start utterance, and the task is considered as successful when the predicted topic is similar enough to the target. We can see that the Kernel agent (Tang et al., 2019) employs single-turn keyword prediction and utilizes the rule strategy to make the topic predicted at each turn is strictly closer to the target than topics selected in preceding turns. As a result, the topics predicted in adjacent turns are distantly related and leading to poor smooth transition. The DKRN agent (Qin et al., 2020) improves the Kernel agent by considering the relations between candidate keywords, and leads to better smooth transition. However, the method still suffers from predicting sub-optimal topics as it overlooks topics in next several turns. For example, in the conversation produced by DKRN, the top-

topic/score	Agent: Kernel Target: car
	A: Hi, stranger, how are you doing? U: Pretty good. Listening to some music.
sound/0.60	A: That sounds relaxing. I am about to settle down and read some poetry. U: What book? I hate reading.
truck/0.80	A: Well, former truck driver, like my dad, but now I just fix them. U: I've never been an artist type.
car/1.0	A: Me too, I also love to race cars.
	Agent: DKRN Target: car
	A: Hi, stranger, how are you doing? U: A bit stressed but alright.
play/0.56	A: I understand that. I got back from playing piano at my father's funeral. U: I am watching my favorite show.
game/0.66	A: I enjoy bingo, we have a game once a week. U: My hobby is watching movie.
music/0.69	A: Wonderful to spend time listening music and away from writing poems. U: Awesome, I also enjoy writing.
car/1.0	A: That is cool. The car is leased anyway so it isn't that impressive.
	Agent: MM-TG Target: car
	A: Hi, stranger, how are you doing? U: I feel a little uncomfortable, want to sleep.
sick/0.53	A: Sorry to hear, are you sick? U: Yes, I am just recovering from an asthma attack and cannot go out.
help/0.66	A: Sorry to hear that, does someone come to help you? U: Yes, luckily my friend found that I was ill.
car/1.0	A: Did he take you to the hospital by car?

Figure 5: Example conversations between user (U) and three different agents (A) with the same target topic (car) and start utterance.

ics are smoothly transitioned in first four turns but changed sharply from "music" to "car" in the fifth turn in order to achieve the target. In contrast, our MM-TP takes a smooth transition from the start topic to the target as "the stranger is *sick* and his friend *helps* him to hospital by *car*". We contribute this to that MM-TP searches the topic thread that will be selected in next several turns by MCTS and then makes prediction.

4.6 Error Analysis

Besides conducting self-play simulation evaluation and performing qualitative evaluation by case study, we also investigate some typical mistakes made by our MM-TP. For example, in Figure 6, the agent communicates with the user for eight turns and finally fail to achieve the target. We attribute this to that the model lacks some common sense knowledge (e.g., knowledge graph of topic) and attempt

topic/score	Agent: MM-TG Target: Baseball
	A: Hi, how are you today? U: Fine. Just finishing up my homework.
reading/0.13	A: Oh, Do you like reading ? U: Yes, I enjoy reading romantic stories.
writing/0.15	A: That sounds helpful for writing U: Yeah, but I do not good at writing articles
singing/0.16	A: Oh cool! I really like music and singing. U: That's great. My favorite singer is Taylor Swift.
dancing/0.24	A: I also love dancing, although I cannot dance well. U: Music is life. Dancing to anything with a beat is the best.
riding/0.26	A: Fine. Sometimes I also like riding in the countryside. U: That sounds relaxing. I love the outdoors.
swimming/0.29	A: Do you like swimming? U: I often walk along a long black river, but I cannot swim.
playing/0.55	A: I also like playing football on weekends. U: I hope I could paly with you.

Figure 6: Example errors made by MM-TG.

to select a near-synonym for the next turn. Specifically, according to the definition of Reward in MM-TP, the transition between topics of consecutive turns should satisfies *smoothness transition* and *target similarity*. However, as not any common sense knowledge are injected into our model, the search policy of MCTS is just trained to select a topic similar to that in the previous turn and more closer to the target. Whether the selected topic is logically related to the topic in the previous turn and can leading the topic thread to the target is overlooked.

5 Related Work

Existing research of dialogue system can be broadly concluded as two categories, which are task-oriented dialogue systems and open-domain dialogue systems. Task-oriented dialogue system aims to accomplish some pre-defined goals (Lipton et al., 2018), conduct negotiation (Cao et al., 2018) or perform symmetric collaborations (He et al., 2017). Open-domain dialogue systems are designed to chat naturally with human and aiming to provide reasonable responses. Previous work make efforts to improve response generation by developing novel neural networks and training on large-scale corpus (Serban et al., 2017; Zhou et al., 2016, 2018). Although the promising progresses have been achieved, these chat-oriented dialogue systems still struggle to a set of limitations such as dull or inconsistent responses (Ram et al., 2018).

Due to these limitations, a novel task named target-guided open-domain conversaion was proposed, which requires the system to proactively and naturally guide the topic thread by integrating goals and strategies. Tang et al. (2019) for the first time introduced this task and employed a simple target-

guided strategy to attain smooth topic transition by turn-level supervised learning. Qin et al. (2020) further improved this work by capturing semantic or factual knowledge relations among candidate topics through a dynamic knowledge routing network. However, both these methods employ single-turn supervised learning to predict the topic of each turn according to the human annotated topic sequence. Moreover, they only consider existing context and overlook the long-term planning of topics in next several turns.

Monte Carlo Tree Search (MCTS) enhanced MDP was firstly proposed in games (Silver et al., 2016; Schrittwieser et al., 2019; Silver et al., 2017) and has been applied in other fields such as diverse ranking (Feng et al., 2018), name entity recognition (Lao et al., 2019) and task-oriented conversation (Wang et al., 2020). In this paper, we apply MCTS in open-domain conversation to generate topic sequence which is utilized to guided the conversation thread to achieve the target.

6 Conclusion

In this paper, we formulate the target-guided conversation as a multi-turn topic prediction problem, and propose a novel approach called MM-TP to resolve this task. MM-TP formalizes the multi-turn topic prediction as sequential decision prediction problem, and models it with MDP. MCTS is used to improve the raw policy by making a long-term planning of topics in next several turns and then selecting a topic for the current turn. The model parameters are learned by reinforcement learning. Experimental results demonstrate that MM-TP outperformed existing baseline systems in terms of both the successful rate of achieving target and the topic transition smoothness.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61702047) and BUPT Excellent Ph.D. Students Foundation (No.CX2020305).

References

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z. Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Demonstrations*, pages 96–100. Association for Computational Linguistics.

C Fellbaum and G Miller. 1998. *WordNet : an electronic lexical database*. MIT Press.

Yue Feng, Jun Xu, Yanyan Lan, Jiafeng Guo, Wei Zeng, and Xueqi Cheng. 2018. From greedy selection to exploratory decision-making: Diverse ranking with policy-value networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 125–134. ACM.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1766–1776. Association for Computational Linguistics.

Zhiting Hu, Haoran Shi, Bowen Tan, Wentao Wang, Zichao Yang, Tiancheng Zhao, Junxian He, Lianhui Qin, Di Wang, Xuezhe Ma, Zhengzhong Liu, Xiaodan Liang, Wanrong Zhu, Devendra Singh Sachan, and Eric P. Xing. 2019. Texar: A modularized, versatile, and extensible toolkit for text generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 159–164. Association for Computational Linguistics.

Yadi Lao, Jun Xu, Sheng Gao, Jun Guo, and Jirong Wen. 2019. Name entity recognition with policy-value networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1245–1248. ACM.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2:*

Short Papers, pages 138–143. Association for Computational Linguistics.

Zachary C. Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2018. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5237–5244. AAAI Press.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1489–1498. Association for Computational Linguistics.

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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Jinghui Qin, Zheng Ye, Jianheng Tang, and Xiaodan Liang. 2020. Dynamic knowledge routing network for target-guided open-domain conversation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8657–8664. AAAI Press.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5427–5436. Association for Computational Linguistics.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: the science behind the alexa prize. *CoRR*, abs/1801.03604.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. 2019. Mastering atari, go, chess and shogi by planning with a learned model. *CoRR*, abs/1911.08265.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. Otters: One-turn topic transitions for open-domain dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2492–2504. Association for Computational Linguistics.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nat.*, 529(7587):484–489.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of go without human knowledge. *Nat.*, 550(7676):354–359.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5624–5634. Association for Computational Linguistics.

Sihan Wang, Kaijie Zhou, Kunfeng Lai, and Jianping Shen. 2020. Task-completion dialogue policy learning via monte carlo tree search with dueling network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3461–3471. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 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 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 496–505. Association for Computational Linguistics.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 65–75. Association for Computational Linguistics.

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 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2020. Keyword-guided neural conversational model. *CoRR*, abs/2012.08383.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 372–381. The Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1118–1127. Association for Computational Linguistics.