arXiv:2006.16779v1 [cs.CL] 30 Jun 2020

# PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning

**Siqi Bao*** **Huang He*** **Fan Wang*** **Hua Wu*** **Haifeng Wang**
**Wenquan Wu** **Zhen Guo** **Zhibin Liu** **Xinchao Xu**
Baidu Inc., China
`{baosiqi, hehuang, wang.fan, wu_hua}@baidu.com`

## Abstract

To build a high-quality open-domain chatbot, we introduce the effective training process of PLATO-2 via curriculum learning. There are two stages involved in the learning process. In the first stage, a coarse-grained generation model is trained to learn response generation under the simplified framework of one-to-one mapping. In the second stage, a fine-grained generation model and an evaluation model are further trained to learn diverse response generation and response coherence estimation, respectively. PLATO-2 was trained on both Chinese and English data, whose effectiveness and superiority are verified through comprehensive evaluations, achieving new state-of-the-art results.

## 1 Introduction

Recently, task agnostic pre-training with large-scale transformer models (Vaswani et al., 2017) and general text corpora has achieved great success in natural language understanding (Devlin et al., 2019) as well as natural language generation, especially open-domain dialogue generation. For instance, based on the general language model GPT-2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020) is further trained for response generation using Reddit comments. To obtain a human-like open-domain chatbot, Meena (Adiwardana et al., 2020) scales up the network parameters to 2.6B and employs more social media conversations in the training process, leading to significant improvement on response quality. To mitigate undesirable toxic or bias traits of large corpora, Blender (Roller et al., 2020) further fine-tunes the pre-trained model with human annotated datasets and emphasizes desirable conversational skills of engagingness, knowledge, empathy and personality.

Besides the above attempts from model scale and data selection, PLATO (Bao et al., 2020) aims to tackle the inherent one-to-many mapping problem to improve response quality. The one-to-many mapping refers to that one dialogue context might correspond to multiple appropriate responses. It is widely recognized that the capability of modeling one-to-many relationship is crucial for response generation (Zhao et al., 2017; Chen et al., 2019). PLATO explicitly models this one-to-many relationship via discrete latent variables, aiming to boost the quality of dialogue generation.

In this work, we will try to scale up PLATO to PLATO-2 and discuss its effective training schema via curriculum learning (Bengio et al., 2009). There are two stages involved in the whole learning process, as sketched in Figure 1. In the first stage, under the simplified one-to-one mapping modeling, a **coarse-grained** generation model is trained for appropriate response generation under different conversation contexts. The second stage continues to refine the generation with a **fine-grained** generation model and an evaluation model. The **fine-grained** generation model explicitly models the one-to-many mapping relationship for diverse response generation. To select the most appropriate responses generated by the fine-grained generation model, the evaluation model is trained to estimate the coherence of the responses.

As for response selection, previous studies have employed variant scoring functions, including forward response generation probability (Adiwardana et al., 2020), backward context recover probability (Zhang et al., 2020) and bi-directional coherence probability (Bao et al., 2020). However, the forward score favors safe and generic responses due to the property of maximum likelihood, while the backward score tends to select the response with a high overlap with the context, resulting in repetitive conversations. In order to ameliorate the above

---

*Equal contribution.

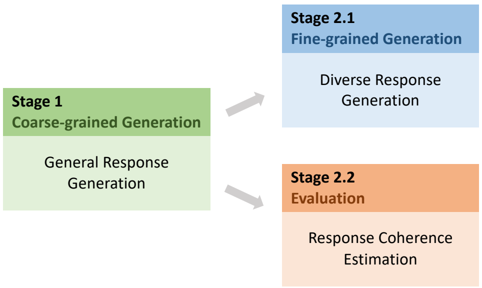

Stage 1
Coarse-grained Generation
General Response Generation
Stage 2.1
Fine-grained Generation
Diverse Response Generation
Stage 2.2
Evaluation
Response Coherence Estimation

Figure 1: Curriculum learning process in PLATO-2.

problems, we adopt the bi-directional coherence estimation in the evaluation model of PLATO-2, whose effectiveness is also verified in the experiments.

We trained PLATO-2 models with different sizes: 1.6 Billion parameters and 310 Million parameters. In addition to the English models, we also trained Chinese models with massive social media conversations. Comprehensive experiments on both English and Chinese datasets demonstrate that PLATO-2 outperforms the state-of-the-art models. We will release our models and source codes at GitHub soon, hoping to facilitate the research in open-domain dialogue generation.[1]

## 2 Methodology

### 2.1 Model Architecture

The infrastructure of PLATO-2 is shown in Figure 2(a), consisted of transformer blocks. The key components of the transform block include layer normalization, multi-head attention and feed forward layers. In the arrangement of these components, there are two options with regarding to the location of layer normalization. One way is the original post-normalization used in BERT (Devlin et al., 2019), where layer normalization is placed between residual connections. The other way is the recent pre-normalization used in GPT-2 (Radford et al., 2019), where layer normalization is placed within residual connections. As reported in Megatron-LM (Shoeybi et al., 2019), post-normalization leads to performance degradation as the model size increases, and pre-normalization enables stable training in large-scale models. As such, pre-normalization is adopted in our model.

Besides, unlike conventional Seq2Seq, there are no separate encoder and decoder networks in our infrastructure. For the sake of training efficiency, PLATO-2 keeps the unified network for bi-directional context encoding and uni-directional response generation through flexible attention mechanism (Dong et al., 2019; Bao et al., 2020).

### 2.2 Curriculum Learning

In this work, we carry out effective training of PLATO-2 via curriculum learning. As shown in Figure 2(b), there are two stages involved in the learning process: during stage 1, a coarse-grained baseline model is trained for general response generation under the simplified one-to-one mapping relationship; during stage 2, two models of fine-grained generation and evaluation are further trained for diverse response generation and response coherence estimation respectively. Although the backbones of these models are all transformer blocks, the self-attention masks are designed accordingly in order to fit the training objectives.

#### 2.2.1 General Response Generation

It is well known that there exists a one-to-many relationship in conversations, where a piece of context may have multiple appropriate responses. Since the classical generation network is designed to fit one-to-one mapping, they tend to generate generic and dull responses. Whereas, it is still an efficient way to capture the general characteristics of response generation. As such, we first train a coarse-grained baseline model to learn general response generation under the simplified relationship of one-to-one mapping. Given one training sample of context and response $(c, r)$, we need to minimize the following negative log-likelihood (NLL) loss:

$$\begin{aligned} \mathcal{L}_{NLL}^{\text{Baseline}} &= -\mathbb{E} \log p(r|c) \\ &= -\mathbb{E} \sum_{t=1}^{T} \log p(r_t|c, r_{<t}) , \end{aligned} \qquad (1)$$

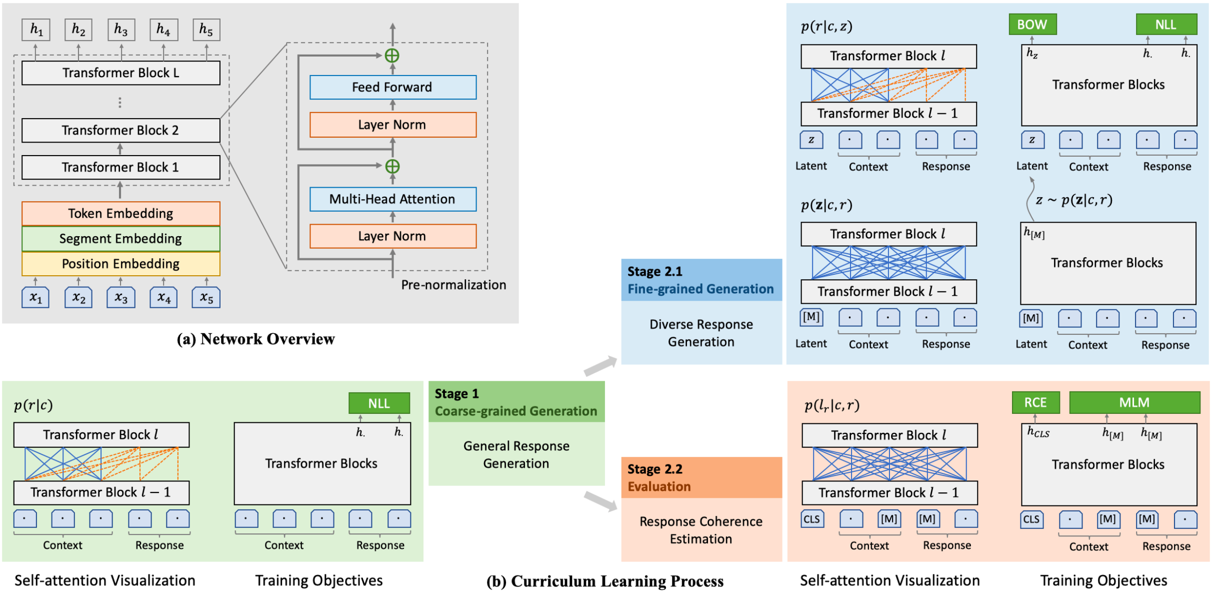
where $T$ is the length of the target response $r$ and $r_{<t}$ denotes previously generated words. Since the response generation is a uni-directional decoding process, each token in the response only attends to those before it, shown as dashed orange lines in Figure 2. As for the context, bi-directional attention is enabled for better natural language understanding, shown as blue lines in Figure 2.

#### 2.2.2 Diverse Response Generation

Based upon the coarse-grained baseline model, diverse response generation is further trained under

---

[1] `https://github.com/PaddlePaddle/plato/plato-2`


$h_1$
$h_2$
$h_3$
$h_4$
$h_5$
Transformer Block L
Transformer Block 2
Transformer Block 1
Token Embedding
Segment Embedding
Position Embedding
$x_1$
$x_2$
$x_3$
$x_4$
$x_5$
Feed Forward
Layer Norm
Multi-Head Attention
Layer Norm
Pre-normalization
$p(r|c,z)$
Transformer Block $l$
Transformer Block $l-1$
BOW
NLL
$h_z$
Transformer Blocks
Latent
Context
Response
$p(\mathbf{z}|c,r)$
$z \sim p(\mathbf{z}|c,r)$
$h_{[M]}$
[M]
Stage 2.1
Fine-grained Generation
Diverse Response Generation
Stage 1
Coarse-grained Generation
General Response Generation
Stage 2.2
Evaluation
Response Coherence Estimation
$p(r|c)$
$p(l_r|c,r)$
RCE
MLM
$h_{CLS}$
CLS
Self-attention Visualization
Training Objectives

(a) Network Overview

(b) Curriculum Learning Process

Figure 2: PLATO-2 illustration. (a) Network overview with the details of transformer blocks. (b) Curriculum learning process with self-attention visualization and training objectives.

the relationship of one-to-many mapping. Following our previous work PLATO (Bao et al., 2020), the discrete latent variable is introduced for the one-to-many relationship modeling. It will first estimate the latent act distribution of the training sample $p(\mathbf{z}|c, r)$ and then generate the response with the sampled latent variable $p(r|c, z)$. The NLL loss of diverse response generation is defined as follows:

$$\begin{aligned}\mathcal{L}_{NLL}^{\text{Generation}} &= -\mathbb{E}_{z\sim p(\mathbf{z}|c,r)} \log p(r|c,z) \\ &= -\mathbb{E}_{z\sim p(\mathbf{z}|c,r)} \sum_{t=1}^{T} \log p(r_t|c,z,r_{<t}) \ ,\end{aligned} \tag{2}$$

where $z$ is the latent act sampled from $p(\mathbf{z}|c, r)$. The posterior distribution over latent values is estimated through the task of latent act recognition:

$$p(\mathbf{z}|c,r) = \text{softmax}(W_1 h_{[M]} + b_1) \in \mathbb{R}^K \ , \tag{3}$$

where $h_{[M]} \in \mathbb{R}^D$ is the final hidden state of the special mask [M], $W_1 \in \mathbb{R}^{K\times D}$ and $b_1 \in \mathbb{R}^K$ denote the weight matrices of one fully-connected layer.

Besides the classical NLL loss, the bag-of-words (BOW) loss (Zhao et al., 2017) is also employed to facilitate the training process of discrete latent variables:

$$\begin{aligned}\mathcal{L}_{BOW}^{\text{Generation}} &= -\mathbb{E}_{z\sim p(\mathbf{z}|c,r)} \sum_{t=1}^{T} \log p(r_t|c,z) \\ &= -\mathbb{E}_{z\sim p(\mathbf{z}|c,r)} \sum_{t=1}^{T} \log \frac{e^{f_{r_t}}}{\sum_{v\in V} e^{f_v}} \ ,\end{aligned} \tag{4}$$

where $V$ refers to the whole vocabulary. The function $f$ tries to predict the words within the target response in a non-autoregressive way:

$$f = \text{softmax}(W_2 h_z + b_2) \in \mathbb{R}^{|V|} \ , \tag{5}$$

where $h_z$ is the final hidden state of the latent variable and $|V|$ is the vocabulary size. $f_{r_t}$ denotes the estimated probability of word $r_t$. As compared with NLL loss, the BOW loss discards the order of words and forces the latent variable to capture the global information of the target response.

To sum up, the objective of the fine-grained generation model is to minimize the following integrated loss:

$$\mathcal{L}^{\text{Generation}} = \mathcal{L}_{NLL}^{\text{Generation}} + \mathcal{L}_{BOW}^{\text{Generation}} \tag{6}$$

### 2.2.3 Response Coherence Estimation

Recently one strategy has been shown effective in boosting response quality, which first generates multiple candidate responses and then performs ranking according to a score function. While there are different definitions about this score function, which can be divided into three categories. Firstly, the length-average log-likelihood is employed in Meena (Adiwardana et al., 2020) as the score function, which considers the forward generation probability of the response given the context $p(r|c)$. Secondly, the maximum mutual information is utilized in DialoGPT (Zhang et al., 2020), which considers the backward probability to recover the context given the candidate response $p(c|r)$. Thirdly, a discriminative function is used in PLATO (Bao et al., 2020) for coherence estimation between the context and response $p(l_r|c, r)$, where $l_r$ is the coherence label. Given that the forward score favors safe responses and the backward score produces repetitive conversations, PLATO-2 adopts the bi-directional coherence estimation in the evaluation model.

The loss of response coherence estimation (RCE) is defined as follows:

$$\begin{aligned}\mathcal{L}_{RCE}^{\text{Evaluation}} = &- \log p(l_r = 1|c, r) \\ &- \log p(l_{r^-} = 0|c, r^-)\end{aligned} \tag{7}$$

The positive training samples come from the dialogue context and corresponding target response $(c, r)$, with coherence label $l_r = 1$. And the negative samples are created by randomly selecting responses from the corpus $(c, r^-)$, with coherence label $l_{r^-} = 0$.

To maintain the capacity of distributed representation, the task of masked language model (MLM) (Devlin et al., 2019) is also included in the evaluation network. Within this task, 15% of the input tokens will be masked at random and the network needs to recover the masked ones. The MLM loss is defined as:

$$\mathcal{L}_{MLM}^{\text{Evaluation}} = -\mathbb{E} \sum_{m \in M} \log p(x_m|x_{\backslash M}), \tag{8}$$

where $x$ refers to the input tokens of context and response. $\{x_m\}_{m \in M}$ stands for masked tokens and $x_{\backslash M}$ denotes the rest unmasked ones.

To sum up, the objective of the evaluation model is to minimize the following integrated loss:

$$\mathcal{L}^{\text{Evaluation}} = \mathcal{L}_{RCE}^{\text{Evaluation}} + \mathcal{L}_{MLM}^{\text{Evaluation}} \tag{9}$$

## 3 Experiments

### 3.1 Training Data

PLATO-2 has English and Chinese models, with training data extracted from open-domain social media conversations. The details are elaborated as follows.

#### 3.1.1 English Data

The English training data is extracted from Reddit comments, which are collected by a third party and made publicly available on pushshift.io (Baumgartner et al., 2020). As the comments are formatted in message trees, any conversation path from the root to a tree node can be treated as one training sample, with the node as response and its former turns as context. To improve the generation quality, we carry out elaborate data cleaning. A message node and its sub-trees will be removed if any of the following conditions is met.

1) The number of BPE tokens is more than 128 or less than 2.
2) Any word has more than 30 characters or the message has more than 1024 characters.
3) The percentage of alphabetic characters is less than 70%.
4) The message contains URL.
5) The message contains special strings, such as r/, u/, &amp.
6) The message has a high overlap with the parent's text.
7) The message is repeated more than 100 times.
8) The message contains offensive words.
9) The subreddit is quarantined.
10) The author is a known bot.

After filtering, the data is split into training and validation sets in chronological order. The training set contains 684M (context, response) samples, ranging from December 2005 to July 2019. For the validation set, 0.2M samples are selected from the rest data after July 2019. The English vocabulary contains 8K BPE tokens (Sennrich et al., 2016), constructed with SentencePiece library.

#### 3.1.2 Chinese Data

The Chinese training data is collected from public domain social medias, followed by a similar cleaning process. After filtering, there are 1.2B (context, response) samples in the training set and 0.1M samples in the validation set. As for the Chinese vocabulary, it contains 30K BPE tokens.

### 3.2 Training Details

PLATO-2 has two model sizes: a small version of 310M parameters and a standard version of 1.6B parameters. The 310M parameter model has 24 transformer blocks and 16 attention heads, with the embedding dimension of 1024. The 1.6B parameter model has 32 transformer blocks and 32 attention heads, with the embedding dimension of 2048.

The hyper-parameters used in the training process are listed as follows. The maximum sequence lengths of context and response are all set to 128. We use Adam (Kingma and Ba, 2015) as the optimizer, with a learning rate scheduler including a linear warmup and an invsqrt decay (Vaswani et al., 2017). To train the large-scale model with a relatively large batch size, we employ gradient checkpointing (Chen et al., 2016) to trade computation for memory. Detailed configurations are summarized in Table 1. The training was carried out on 64 Nvidia Tesla V100 32G GPU cards. It takes about 3 weeks for the 1.6B parameter model to accomplish the curriculum learning process.

## 3.3 Evaluation Settings

### 3.3.1 Compared Methods

The following methods have been compared in the experiments.

- DialoGPT (Zhang et al., 2020) is trained on the basis of GPT-2 (Radford et al., 2019) using Reddit comments. There are three model sizes: 117M, 345M and 762M. Since the 345M parameter model obtains the best performance in their evaluations, we compare with it in the experiments.
- Blender (Roller et al., 2020) is firstly trained using Reddit comments and then fine-tuned with human annotated conversations, to help emphasize desirable conversational skills of engagingness, knowledge, empathy and personality. During fine-tuning, there are four datasets included: ConvAI2 (Zhang et al., 2018; Dinan et al., 2020), Empathetic Dialogues (Rashkin et al., 2019), Wizard of Wikipedia (Dinan et al., 2019) and Blended Skill Talk (Smith et al., 2020). These annotated conversations are referred as BST in short. Blender has three model sizes: 90M, 2.7B and 9.4B. Since the 2.7B parameter model obtains the best performance in their evaluations, we compare with it in the experiments.
- Meena (Adiwardana et al., 2020) is an open-domain chatbot trained with social media conversations. Meena has 2.6B model parameters, similar to Blender. Given that Meena has not released the model or provided a service interface, it is difficult to perform comprehensive comparison. In the experiments, we include the provided samples in their paper for static evaluation.
- Microsoft XiaoIce (Zhou et al., 2020) is a popular social chatbot in Chinese. In the experiments, we use the official Weibo platform to chat with XiaoIce.

For the sake of comprehensive and fair comparisons, three versions of PLATO-2 are included in the experiments.

- PLATO 1.6B parameter model is the standard version in English, which is first trained using Reddit comments and then fine-tuned with BST conversations. This model will be compared to the state-of-the-art open-domain chatbot Blender, to measure the effectiveness of PLATO-2.
- PLATO 310M parameter model is a small version in English, which is trained with Reddit comments. This model will be compared to DialoGPT, as they have similar model scales and training data.
- PLATO 333M parameter Chinese model[2] will be compared to XiaoIce in the experiments.

### 3.3.2 Evaluation Metrics

We carry out both automatic and human evaluations in the experiments. In automatic evaluation, to assess the model's capacity on lexical diversity, we use the corpus-level metric of distinct-1/2 (Li et al., 2016a), which is defined as the number of distinct uni- or bi-grams divided by the total number of generated words.

In human evaluation, we employ four utterance-level and dialogue-level metrics, including coherence, informativeness, engagingness and humanness. Three crowd-sourcing workers are asked to score the response/dialogue quality on a scale of [0, 1, 2], with the final score determined through majority voting. The higher score, the better. These criteria are discussed as follows, with scoring details provided in the Appendix.

- Coherence is an utterance-level metric, measuring whether the response is relevant and consistent with the context.
- Informativeness is also an utterance-level metric, evaluating whether the response is informative or not given the context.
- Engagingness is a dialogue-level metric, assessing whether the annotator would like to talk with the speaker for a long conversation.
- Humanness is also a dialogue-level metric, judging whether the speaker is a human being or not.

---

[2] This model has 24 transformer blocks and 16 attention heads, with the embedding dimension of 1024. As the Chinese vocabulary contains 30K BPE tokens, this model has 23M more parameters than the English small model.

| Model | Stage | Batch Size | Learning Rate | Warmup Steps | Training Steps |
|---|---|---|---|---|---|
| 310M Parameter | 1. Baseline | 524,288 | 1e-3 | 4000 | 20w |
| | 2.1. Generation | 262,144 | 4e-5 | 4000 | 10w |
| | 2.2. Evaluation | 524,288 | 1e-4 | 4000 | 10w |
| 1.6B Parameter | 1. Baseline | 524,288 | 5e-4 | 3125 | 20w |
| | 2.1. Generation | 65,536 | 1e-5 | 3125 | 10w |
| | 2.2. Evaluation | 524,288 | 5e-5 | 3125 | 10w |

Table 1: Training configurations of PLATO-2.

| Training Data | Model | Human Evaluation | | | | Automatic Evaluation | | Average Length |
|---|---|---|---|---|---|---|---|---|
| | | Coherence | Informativeness | Engagingness | Humanness | Distinct-1 | Distinct-2 | |
| Reddit | DialoGPT | 0.720 | 0.712 | 0.340 | 0.100 | 0.143 | 0.504 | 9.906 |
| | PLATO 310M | 1.572 | 1.620 | 1.300 | 1.160 | 0.064 | 0.432 | 23.536 |
| Reddit + BST | Blender | 1.856 | 1.816 | 1.820 | 1.540 | 0.115 | 0.387 | 17.677 |
| | PLATO 1.6B | **1.920** | **1.892** | **1.840** | **1.740** | **0.163** | **0.603** | 16.540 |

Table 2: Self-chat evaluation results, with best value written in bold.

### 3.4 Experimental Results

In the experiments, we include both static and interactive evaluations.

### 3.4.1 Self-Chat Evaluation

Self-chats have been widely used in the evaluation of dialogue systems (Li et al., 2016b; Bao et al., 2019; Roller et al., 2020), where a model plays the role of both partners in the conversation. As compared with human-bot conversations, self-chat logs can be collected efficiently at a cheaper price. As reported in Li et al. (2019), self-chat evaluations exhibit high agreement with the human-bot chat evaluations. In the experiments, we ask the bot to perform self-chats and then invite crowd-sourcing workers to evaluate the dialogue quality.

The way to start the interactive conversation needs special attention. As pointed out by Roller et al. (2020), if starting with 'Hi!', partners tend to greet with each other and only cover some shallow topics in the short conversation. Therefore, to expose the model's weaknesses and explore the model's limits, we choose to start the interactive conversation with pre-selected topics. We use the classical 200 questions as the start topic (Vinyals and Le, 2015) and ask the bot to performance self-chats given the context. There are 10 utterances in each dialogue, including the input start utterance. We carry out automatic evaluation on the 200 self-chat logs and randomly select 50 conversations from 200 self-chat logs for human evaluation.

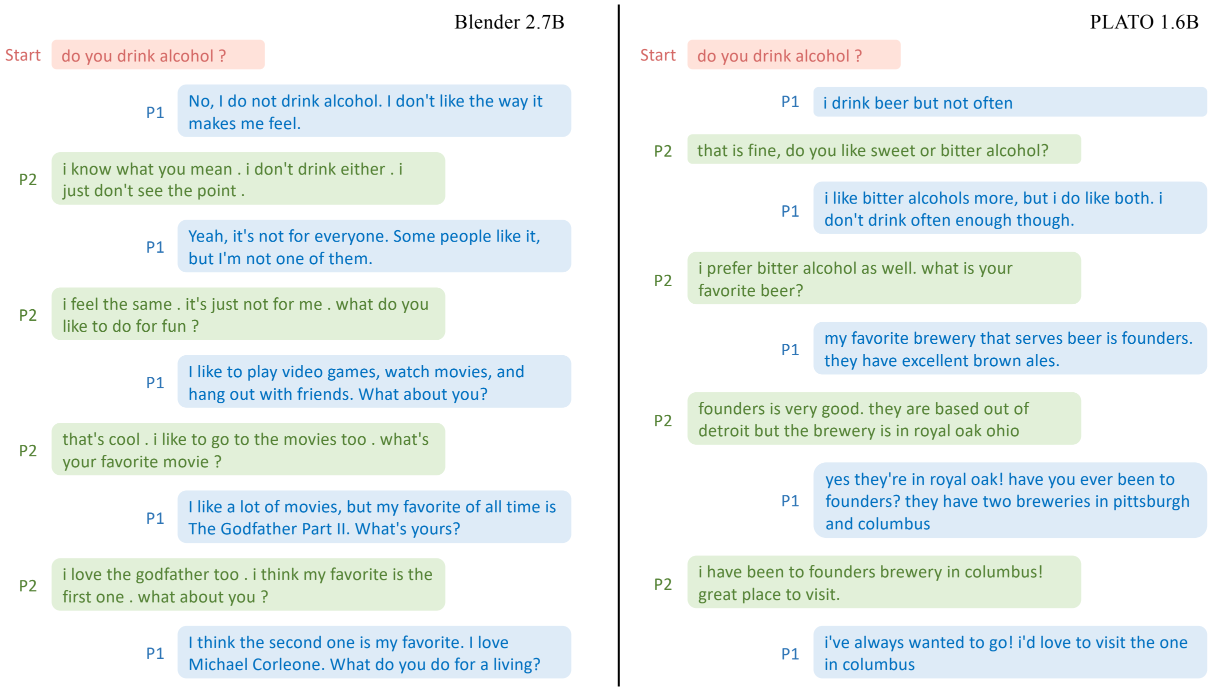
The compared models are divided into two groups. The first group includes DialoGPT 345M model and PLATO 310M model. Both of them are trained using Reddit comments and have similar model scales. The second group includes Blender 2.7B model and PLATO 1.6B model. Both of them are trained using Reddit comments and further fine-tuned with BST conversations. In human evaluation, two self-chat logs, which are from the same group and have the same start topic, will be displayed to three crowd-sourcing workers. One example is given in Figure 3. As suggested in ACUTE-Eval (Li et al., 2019), we ask crowd-sourcing workers to pay attention to only one speaker within a dialogue. In the evaluation, they need to give scores on coherence and informativeness for each P1's utterance, and assess P1's overall quality on engagingness and humanness. The self-chat evaluation results are summarized in Table 2. These results demonstrate that PLATO 1.6B model obtains the best performance across human and automatic evaluations. The gap of Blender and PLATO-2 on the corpus-level metric distinct-1/2 suggests that PLATO-2 has a better capacity on lexical diversity. In addition, the difference between these two groups indicates that enlarging model scales and exploiting human annotated conversations help improve the dialogue quality.

### 3.4.2 Human-Bot Chat Evaluation

In the Chinese evaluation, it is difficult to carry out self-chats for Microsoft XiaoIce, as there is no public available API. Therefore, we collect human-

| Model | Human Evaluation | | | | Automatic Evaluation | | Average Length |
|---|---|---|---|---|---|---|---|
| | Coherence | Informativeness | Engagingness | Humanness | Distinct-1 | Distinct-2 | |
| Microsoft XiaoIce | 0.869 | 0.822 | 0.560 | 0.260 | **0.289** | **0.764** | 6.979 |
| PLATO 333M Chinese | **1.737** | **1.683** | **1.600** | **1.480** | 0.212 | 0.713 | 6.641 |

Table 3: Chinese interactive evaluation results, with best value written in bold.



Blender 2.7B

Start do you drink alcohol ?

P1 No, I do not drink alcohol. I don't like the way it makes me feel.

P2 i know what you mean . i don't drink either . i just don't see the point .

P1 Yeah, it's not for everyone. Some people like it, but I'm not one of them.

P2 i feel the same . it's just not for me . what do you like to do for fun ?

P1 I like to play video games, watch movies, and hang out with friends. What about you?

P2 that's cool . i like to go to the movies too . what's your favorite movie ?

P1 I like a lot of movies, but my favorite of all time is The Godfather Part II. What's yours?

P2 i love the godfather too . i think my favorite is the first one . what about you ?

P1 I think the second one is my favorite. I love Michael Corleone. What do you do for a living?

PLATO 1.6B

Start do you drink alcohol ?

P1 i drink beer but not often

P2 that is fine, do you like sweet or bitter alcohol?

P1 i like bitter alcohols more, but i do like both. i don't drink often enough though.

P2 i prefer bitter alcohol as well. what is your favorite beer?

P1 my favorite brewery that serves beer is founders. they have excellent brown ales.

P2 founders is very good. they are based out of detroit but the brewery is in royal oak ohio

P1 yes they're in royal oak! have you ever been to founders? they have two breweries in pittsburgh and columbus

P2 i have been to founders brewery in columbus! great place to visit.

P1 i've always wanted to go! i'd love to visit the one in columbus

Figure 3: Self-chat examples by Blender and PLATO-2.

bot conversations through their official Weibo platform. The interactive conversation also starts with a pre-selected topic and continues for 7-14 rounds, where 50 diverse topics are extracted from the high-frequency topics of a commercial chatbot, including travel, movie, hobby and so on. The collected human-bot conversations are distributed to crowd-sourcing workers for evaluation. The human and automatic evaluation results are summarized in Table 3. XiaoIce obtains higher distinct values, which may use a retrieval-based strategy in response generation. The human evaluations demonstrate that our PLATO-2 model achieves significant improvements over XiaoIce across all the human evaluation metrics.

### 3.4.3 Static Evaluation

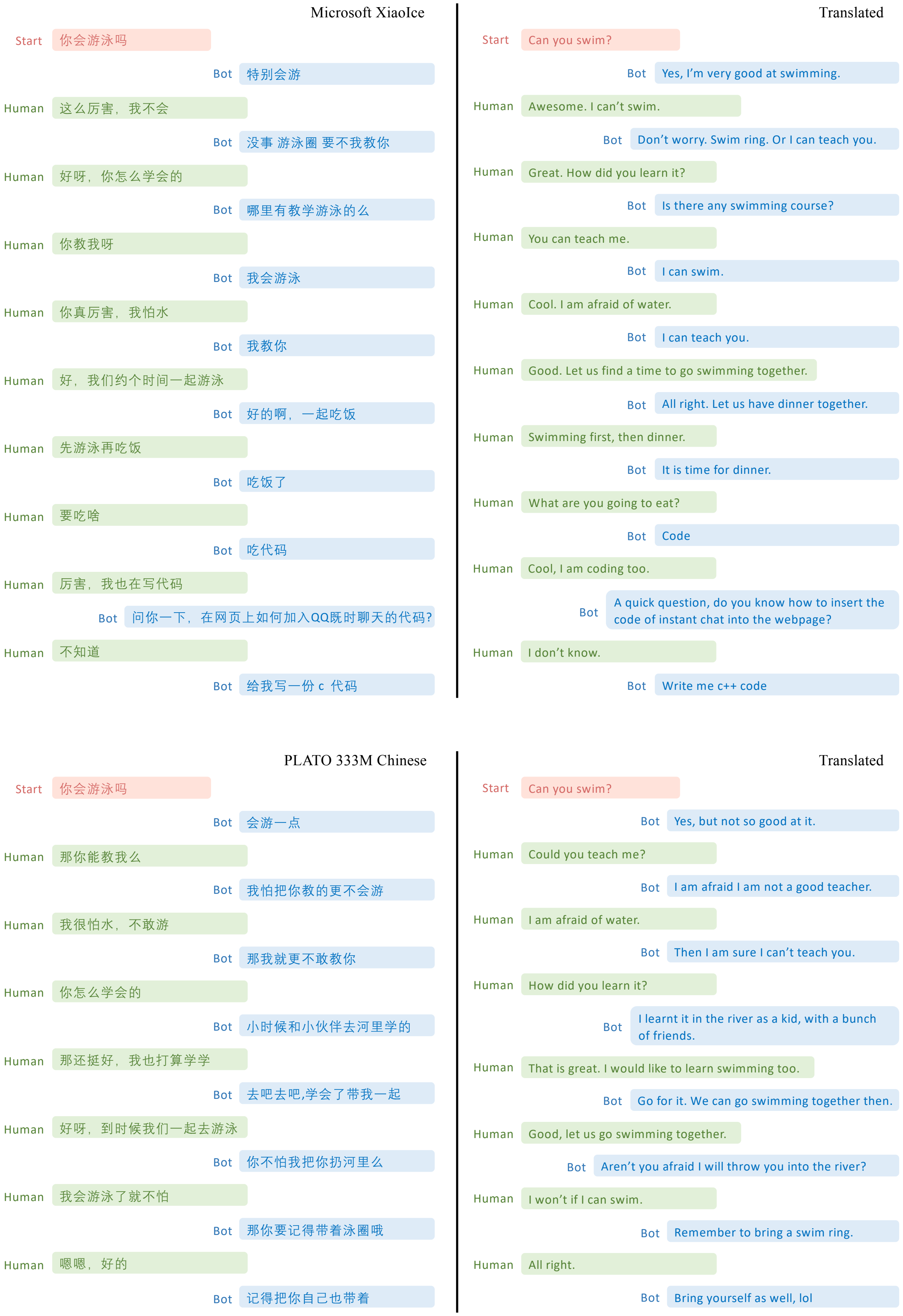
Besides the interactive evaluation, we also include static evaluation to analyze the model's performance. In static evaluation, each model will produce a response towards the given multi-turn context. To compare with Meena, we include their provided 60 static samples in the paper's Appendix

| Model | Meena Samples | | |
|---|---|---|---|
| | Coherence | Informativeness | Engagingness |
| Meena | 1.750 | 1.617 | 1.583 |
| DialoGPT | 1.233 | 1.067 | 1.017 |
| Blender | 1.800 | 1.767 | 1.683 |
| PLATO 1.6B | **1.900** | **1.917** | **1.850** |

| Model | Daily Dialog Samples | | |
|---|---|---|---|
| | Coherence | Informativeness | Engagingness |
| DialoGPT | 1.117 | 1.033 | 0.917 |
| Blender | 1.767 | 1.617 | 1.633 |
| PLATO 1.6B | **1.867** | **1.850** | **1.833** |

| Model | Reddit Samples | | |
|---|---|---|---|
| | Coherence | Informativeness | Engagingness |
| DialoGPT | 1.283 | 1.283 | 1.183 |
| Blender | 1.767 | 1.550 | 1.583 |
| PLATO 1.6B | **1.900** | **1.900** | **1.883** |

Table 4: Static evaluation results, with best value written in bold.



Microsoft XiaoIce

Start 你会游泳吗

Bot 特别会游

Human 这么厉害，我不会

Bot 没事 游泳圈 要不我教你

Human 好呀，你怎么学会的

Bot 哪里有教学游泳的么

Human 你教我呀

Bot 我会游泳

Human 你真厉害，我怕水

Bot 我教你

Human 好，我们约个时间一起游泳

Bot 好的啊，一起吃饭

Human 先游泳再吃饭

Bot 吃饭了

Human 要吃啥

Bot 吃代码

Human 厉害，我也在写代码

Bot 问你一下，在网页上如何加入QQ既时聊天的代码?

Human 不知道

Bot 给我写一份 c 代码

Translated

Start Can you swim?

Bot Yes, I'm very good at swimming.

Human Awesome. I can't swim.

Bot Don't worry. Swim ring. Or I can teach you.

Human Great. How did you learn it?

Bot Is there any swimming course?

Human You can teach me.

Bot I can swim.

Human Cool. I am afraid of water.

Bot I can teach you.

Human Good. Let us find a time to go swimming together.

Bot All right. Let us have dinner together.

Human Swimming first, then dinner.

Bot It is time for dinner.

Human What are you going to eat?

Bot Code

Human Cool, I am coding too.

Bot A quick question, do you know how to insert the code of instant chat into the webpage?

Human I don't know.

Bot Write me c++ code

PLATO 333M Chinese

Start 你会游泳吗

Bot 会游一点

Human 那你能教我么

Bot 我怕把你教的更不会游

Human 我很怕水，不敢游

Bot 那我就更不敢教你

Human 你怎么学会的

Bot 小时候和小伙伴去河里学的

Human 那还挺好，我也打算学学

Bot 去吧去吧,学会了带我一起

Human 好呀，到时候我们一起去游泳

Bot 你不怕我把你扔河里么

Human 我会游泳了就不怕

Bot 那你要记得带着泳圈哦

Human 嗯嗯，好的

Bot 记得把你自己也带着

Translated

Start Can you swim?

Bot Yes, but not so good at it.

Human Could you teach me?

Bot I am afraid I am not a good teacher.

Human I am afraid of water.

Bot Then I am sure I can't teach you.

Human How did you learn it?

Bot I learnt it in the river as a kid, with a bunch of friends.

Human That is great. I would like to learn swimming too.

Bot Go for it. We can go swimming together then.

Human Good, let us go swimming together.

Bot Aren't you afraid I will throw you into the river?

Human I won't if I can swim.

Bot Remember to bring a swim ring.

Human All right.

Bot Bring yourself as well, lol

Figure 4: Human-bot chat examples by Microsoft XiaoIce and PLATO-2.

and generate corresponding responses with other models. We also include 60 test samples about daily life from Daily Dialog (Li et al., 2017) and 60 test samples about in-depth discussion from Reddit. Given that the measurement of humanness usually needs multi-turn interaction, this metric is excluded from static evaluation. The evaluation results are summarized in Table 4. It can be observed that PLATO-2 is able to produce coherent, informative and engaging responses under different chat scenarios.

## 3.5 Discussions

### 3.5.1 Case Analysis

To further analyze the models' features, two self-chat examples of Blender and PLATO-2 are provided in Figure 3. Although both models are able to produce high-quality engaging conversations, they exhibit distinct discourse styles. Blender tends to switch topics quickly in the short conversation, including alcohol, hobbies, movies and work. The emergence of this style might be related with BST fine-tuning data. For instance, ConvAI2 is about the exchange of personal information between two partners, where topics need to switch quickly to know more about each other. Due to the task settings of data collection, some human annotated conversations might be a little unnatural. Nevertheless, fine-tuning with BST conversations is essential to mitigate undesirable toxic traits of large corpora and emphasize desirable skills of human conversations.

Distinct with Blender, PLATO-2 can stick to the start topic and conduct in-depth discussions. The reasons might be two-fold. On the one hand, our model is able to generate diverse and informative responses with the accurate modeling of one-to-many relationship. On the other hand, the evaluation model helps select the coherent response and stick to current topic. We also asked crowd-sourcing workers to compare the responses generated by these two models via pairwise ranking. The comparison result is shown in Table 5, which also verifies our above analysis on discourse styles.

Besides, we also provide two human-bot chat examples of XiaoIce and PLATO-2 in Figure 4, with original interactive logs shown on the left and translated logs on the right. It can be observed that some responses produced by XiaoIce are not coherent with the contexts and there are some abrupt changes of topics. By contrast, the interaction with PLATO-2 is more coherent and engaging.

| Blender Win | Tie | PLATO 1.6B Win |
|---|---|---|
| 8 | 18 | **24** |

Table 5: Thoroughness with regard to the start topic.

| Score Function | MAP | MRR | P@1 |
|---|---|---|---|
| $p(r\|c)$ | 0.705 | 0.790 | 0.700 |
| $p(c\|r)$ | 0.672 | 0.737 | 0.610 |
| $p(l_r\|c,r)$ | **0.754** | **0.819** | **0.750** |

Table 6: Comparison of different score functions in response selection, with best value written in bold.

### 3.5.2 Response Selection Comparison

We carry out more experiments to compare the performance of distinct score functions in response selection. Firstly, one Chinese response selection dataset is constructed: 100 dialogue contexts are selected from the test set and 10 candidate responses are retrieved for each context with a commercial chatbot. Secondly, we ask crowd-sourcing workers to annotate the label whether the candidate response is coherent with the context. Thirdly, we train three 333M parameter models as the score function, including the forward response generation probability $p(r|c)$, the backward context recover probability $p(c|r)$ and the bi-directional coherence probability $p(l_r|c,r)$. Their results on the annotated response selection dataset are summarized in Table 6. The metrics of mean average precision (MAP) (Baeza-Yates et al., 1999), mean reciprocal rank (MRR) (Voorhees et al., 1999) and precision at position 1 (P@1) are employed. These results indicate that PLATO-2's evaluation model is better at selecting appropriate responses.

## 4 Related Work

Related works include large-scale language models and open-domain dialogue generation.

**Large-scale Language Models.** Pre-trained large-scale language models have brought many breakthroughs on various NLP tasks. GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) are representative uni-directional and bi-directional language models, trained on general text corpora. By introducing pre-normalization and modifying weight initialization, GPT-2 (Radford et al., 2019) successfully extends the model scale from 117M to 1.5B parameters. To cope with memory constraints, Megatron-LM (Shoeybi et al., 2019) exploits model parallelism to train an 8.3B parameter model on 512 GPUs. GPT-3 (Brown et al., 2020) further trains an 175B parameter autoregressive language model, demonstrating strong performance on many NLP tasks. The development of large-scale language models is also beneficial to the task of dialogue generation.

**Open-domain Dialogue Generation.** On the basis of GPT-2, DialoGPT (Zhang et al., 2020) is trained for response generation using Reddit comments. To obtain a human-like open-domain chatbot, Meena (Adiwardana et al., 2020) scales up the network parameters to 2.6B and utilizes more social media conversations in the training process. To emphasize desirable conversational skills of engagingness, knowledge, empathy and personality, Blender (Roller et al., 2020) further fine-tunes the pre-trained model with human annotated conversations. Besides the above attempts on model scale and data selection, PLATO (Bao et al., 2019) introduces discrete latent variable to tackle the inherent one-to-many mapping problem to improve response quality. In this work, we further scale up PLATO to PLATO-2 and discuss its effective training via curriculum learning.

## 5 Conclusion

In this work, we discuss the effective training of open-domain chatbot PLATO-2 via curriculum learning, where two stages are involved. In the first stage, one coarse-grained model is trained for general response generation. In the second stage, two models of fine-grained generation and evaluation are trained for diverse response generation and response coherence estimation. Experimental results demonstrate that PLATO-2 achieves substantial improvements over the state-of-the-art methods in both Chinese and English evaluations.

## Acknowledgments

We would like to thank Jingzhou He, Tingting Li, Daxiang Dong and Pingshuo Ma for their generous help. This work was supported by the Natural Key Research and Development Project of China (No. 2018AAA0101900).

## References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Siqi Bao, Huang He, Fan Wang, Rongzhong Lian, and Hua Wu. 2019. Know more about each other: Evolving dialogue strategy via compound assessment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5382–5391.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4918–4924.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *International Conference on Learning Representations*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 986–995.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report, OpenAI*.

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*, pages 5370–5381.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

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*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill 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: System Demonstrations*, pages 270–278.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 654–664.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

# A Scoring Criteria

| Score | Coherence |
|---|---|
| 0 | • The response is not related with the context.<br>• The response simply repeats the context.<br>• The response has obvious conflicts with the context.<br>• There are serious logic conflicts within the response. |
| 1 | • The response has minor conflicts with the context.<br>• There are some minor logic conflicts in the response. |
| 2 | • The response is coherent with the context. |

| Score | Informativeness |
|---|---|
| 0 | • The response doesn't contain any information.<br>• This response just repeats the context and fails to bring any additional information.<br>• The information is invalid, as the coherence score is 0. |
| 1 | • The information has conflicts with common sense.<br>• There are factual errors in the response. |
| 2 | • The response has appropriate and correct information. |

| Score | Engagingness |
|---|---|
| 0 | • I don't want to talk with this speaker. |
| 1 | • It is kind of boring, but it is still ok to talk with this speaker. |
| 2 | • I would like to talk with this speaker for a long conversation. |

| Score | Humanness |
|---|---|
| 0 | • This speaker seems like a bot. |
| 1 | • This speaker gives unnatural responses occasionally and seems not that human-like. |
| 2 | • This speaker seems like a human being. |

Table 7: Score details of four metrics in human evaluation.