A Simple and Efficient Multi-Task Learning Approach for Conditioned Dialogue Generation

Yan Zeng
DIRO, Université de Montréal
`yan.zeng@umontreal.ca`

Jian-Yun Nie
DIRO, Université de Montréal
`nie@iro.umontreal.ca`

Abstract

Conditioned dialogue generation suffers from the scarcity of labeled responses. In this work, we exploit labeled non-dialogue text data related to the condition, which are much easier to collect. We propose a multi-task learning approach to leverage both labeled dialogue and text data. The 3 tasks jointly optimize the same pre-trained Transformer – conditioned dialogue generation task on the labeled dialogue data, conditioned language encoding task and conditioned language generation task on the labeled text data. Experimental results show that our approach outperforms the state-of-the-art models by leveraging the labeled texts, and it also obtains larger improvement in performance comparing to the previous methods to leverage text data.

1 Introduction

General conversational models pre-trained on large text data (Radford et al., 2018; Devlin et al., 2018) or human-to-human conversation data (Zhang et al., 2019; Bao et al., 2019) have shown excellent performance in generating fluent and diverse responses. In addition to general conversation, we are more and more faced with the problem of conditioned conversation that tunes the dialogue toward a specific style or domain. For example, we might specify a condition as the vocabulary frequently used by a person and ask the system to mimic the speaking style of the person, or a topic-related vocabulary and ask the chatbot to discuss the given topic.

Conditioned response generation has been extensively explored using RNN-based sequence-to-sequence models, under different conditions, e.g. persona (Li et al., 2016b), topic (Xing et al., 2017), emotion (Zhou et al., 2018), situations (Sato et al., 2017), and so on. However, only a few existing studies considered using pre-training based models (Zheng et al., 2019; Lin et al., 2019). The basic idea in these previous works is to utilize a parametric vector to represent a condition and then use it in

the decoder for conditioned generation. However, the key issue in conditioned dialogue generation is the availability of labeled responses (Zhou and Wang, 2018), and pre-training on unlabeled text or dialogue data does not help much. Therefore, the motivation of this work is to leverage labeled text (non-dialogue) data that are much easier to collect than labeled dialogue data as supplement. These data can be, for example, texts written by the same person (for a persona condition), within the same topic domain (for a topic condition), etc. The idea is inspired by response style transfer (Luan et al., 2017; Niu and Bansal, 2018), which uses a text corpus to learn a style and then transfer the style to dialogue. Based on their success, we assume that the labeled text data can contribute to create better representations of conditions and better utilization of conditions in natural language generation.

In this work, we propose a multi-task learning approach to leverage both labeled dialogue and text data. We use 3 tasks to jointly optimize the same pre-trained Transformer – conditioned dialogue generation task on the labeled dialogue data, conditioned language encoding task and conditioned language generation task on the labeled text data. Our assumption is that the two other tasks can help in our final goal of conditioned dialogue generation: conditioned language generation is the base of conditioned response generation, and conditioned language encoding using bi-directional attention can efficiently encode condition-related expressions and lead to better condition representations. We apply different input representations, self-attention masks, and random mask strategies to differentiate the 3 tasks. Regardless of these differences, the training objectives of these tasks are essentially the same, i.e. masked language modeling, and thus we can mix up 2 types of data / 3 tasks in one training batch, which prevents us from having the catastrophic forgetting problem (Phang et al., 2018).

To efficiently leverage labeled data, first, our approach incorporates all types of data within the same framework, avoiding introducing ad hoc model components which are usually needed in some response style transfer methods in order to leverage extra texts. Second, we propose *TF-IDF based masking* which selects more condition-related tokens to mask, so that the model can exploit the labeled text data more for condition-related expressions rather than the general language features already captured by the pre-trained models. Third, for conditioned generation, we propose a *non-parametric attention-based gating mechanism*, which chooses between generating a general word (necessary for general function words) or a condition-related word at each position. We expect it to be more efficient than a parametric gating. Experimental results show that these approaches all bring improvements.

Our approach is generalizable. In spite of many different labels, a condition essentially specifies some preferences on words, phrases, and sentence structures in the generated responses. Thus, a general approach can be instantiated to a specific case as long as the corresponding labeled dialogue data are available. We will run experiments with two instantiated models for persona- and topic-related dialogue. Additionally, we will empirically show that our approach is robust and can even work with condition labels predicted by a classification model, e.g. LDA for topic labels.

The contributions in this work are as follows [1]:

- We propose a simple and efficient multi-task learning approach based on pre-trained Transformer that leverages different labeled data, i.e., dialogue and text, for conditioned response generation.

- The experiments under two different conditions – persona- and topic-based dialogue, show that our approach outperforms the state-of-the-art models by leveraging labeled texts even when the labels are predicted by a model.

- Our approach obtains larger improvement in performance comparing to the existing methods to leverage text data, based on extra auto-encoder or sequential fine-tuning.

2 Related Work

2.1 Conditioned Dialogue Generation

We categorize the related existing works into 3 categories. (1) Response generation conditioned on latent variables, where no extra annotations of dialogues is required (Serban et al., 2017; Shen et al., 2018; Gu et al., 2018; Chen et al., 2019; Gao et al., 2019; Bao et al., 2020). (2) Loosely-conditioned response generation, where a label designating the type of the response is required. For example, persona labels (Li et al., 2016b) designate the speaking styles of the responses, and topic labels (Xing et al., 2017; Dziri et al., 2019) or emotion labels (Li et al., 2017; Zhou et al., 2018; Rashkin et al., 2019) specify topic-related or emotion-related vocabularies. These studies usually utilize a parametric vector to encode a label, which is then used in the decoder to guide the generation. (3) Strictly-conditioned response generation, where extra knowledge is required to determine the content of the response, such as a persona profile (Zhang et al., 2018; Urbanek et al., 2019), a situation description (Rashkin et al., 2018; Urbanek et al., 2019), or a wikipedia paragraph (Galley et al., 2019; Dinan et al., 2018), which are used to ground the response. The ability to strictly-conditioned generation is important, but these dialogues only count for a small fraction of open-domain conversation (Zheng et al., 2019). In many other cases, we are in the situation of loosely-conditioned dialogue. Furthermore, the state-of-the-art strictly-conditioned method (Wolf et al., 2019) can be easily added in other models as well (Shuster et al., 2019; Madotto et al., 2020), which simply concatenates the extra knowledge with the dialogue history as the model input.

In this work, we focus on loosely-conditioned response generation [2]. We will show that our approach is robust and can work with different types of labels including those predicted by a classification model, e.g. LDA for topic labels. Therefore, our method is compatible to generation conditioned on latent variables by borrowing power of a classification model. In this work, we do not touch on strictly-conditioned generation. However, this ability can be easily equipped as mentioned.

2.2 Response Style Transfer

Style transfer in dialogue aims to learn the style of a text corpus and then incorporate the style in dia-

---

[2]Conditioned generation elsewhere in this work refers to loosely-conditioned generation.

logue generation. The transfer is usually between two styles, e.g. rude and polite, or adding a style to general dialogues. To leverage the text corpus, Luan et al. (2017) jointly trains a seq2seq response generator and an extra auto-encoder, and Niu and Bansal (2018) trains an extra style classifier first to guild the response generator using reinforcement learning.

These works show that text data contain rich information about how to generate a specific type of texts, which inspire us to exploit the labeled text data in conditioned dialogue generation to alleviate the data scarcity issue. Style transfer is usually between two given styles. In contrast, conditioned dialogue generation could work with hundreds of condition labels simultaneously. As we will show in our experiments, the style transfer methods that utilize additional models, e.g. auto-encoder, to leverage text corpus are unscalable and inefficient for conditioned dialogue. In contrast, our approach that leverages labeled text data without using ad hoc models and makes a tighter integration of labeled text data with labeled dialogue data can more directly impact the conditioned dialogue generation.

3 Method

We assume that we have two types of training data: a labeled dialogue corpus containing (dialogue history, condition, response) samples, and a labeled text corpus consisting of (condition, text) samples. Notice that the "condition" is any categorical label that indicates a type of responses or texts. Our goal is to generate a response y that exhibits the desired characteristics of the type of responses given a dialogue history x and a condition c:

$$y = \arg\max_y P(y|x, c) \qquad (1)$$

The Transformer in our work uses bi-directional attention on the source side to encode the dialogue history, and left-to-right attention on the target side to generate the response. Such a transformer can be initialized from BERT(Devlin et al., 2018), Roberta(Liu et al., 2019), UniLM (Dong et al., 2019), or the models pre-trained on large-scale unlabeled dialogue data e.g. PLATO (Bao et al., 2019) and Blender (Roller et al., 2020). In this work, we focus on efficiently leveraging labeled data, i.e. dialogue and text. Figure 1 (Left) shows the overview of our approach.

3.1 Masked Multi-Head Attention

In this subsection, we introduce the basic components of Transformer. Masked multi-head attention is also applied in our condition-aware transformer block. The input representation $\mathbf{H}^0 \in \mathbb{R}^{n \times d_h}$, where n is the input length and $d_h = 768$ is the hidden dimension, is the sum of token embedding, position embedding, and type embedding at each position. We apply type embeddings to introduce a separation between source side and target side as shown in Figure 1 (Left) in order to warrant different treatments in the model. Then, $\mathbf{H}^0$ is encoded into hidden representations of i-th layer $\mathbf{H}^i = [\mathbf{h}_1^i, ..., \mathbf{h}_n^i]$ using multi-layer transformer blocks: $\mathbf{H}^i = \text{Trans}^i(\mathbf{H}^{i-1}) \quad i \in [1, L]$. The core component of a transformer block is the masked multi-head attention, whose outputs, i.e. contextualized representations, $\mathbf{C}^i = [\mathbf{c}_1^i, ..., \mathbf{c}_n^i]$, are computed via: $\mathbf{C}^i = \text{Concat}(\mathbf{head}_1, ..., \mathbf{head}_h)$. Specifically,

$$\mathbf{head}_j = \text{softmax}(\frac{\mathbf{Q}_j \mathbf{K}_j^T}{\sqrt{d_k}} + \mathbf{M})\mathbf{V}_j \qquad (2)$$

where $\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j \in \mathbb{R}^{n \times d_k}$ are obtained by transforming $\mathbf{H}^{i-1} \in \mathbb{R}^{n \times d_h}$ using $\mathbf{W}_j^{iQ}, \mathbf{W}_j^{iK}, \mathbf{W}_j^{iV} \in \mathbb{R}^{d_h \times d_k}$ respectively. The self-attention mask matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ (with $\mathbf{M}_{ij} \in \{0, -\infty\}$) determines whether a position can attend to other positions: $\mathbf{M}_{ij} = 0$ allows the i-th position to attend to j-th position and $\mathbf{M}_{ij} = -\infty$ prevents from it.

Our approach jointly optimizes three tasks that apply different self-attention masks as shown in Figure 1 (Left). For conditioned dialogue generation task, the self-attention mask allows bi-directional attention on the source side to fully encode dialogue history and left-to-right attention on the target side to generate conditioned responses. For the labeled text data, we randomly choose between conditioned language encoding and conditioned language generation task. The two tasks use bi-directional attention and left-to-right attention respectively. The language encoding objective, i.e. Masked Language Modeling (MLM), is used in BERT, which has shown stronger ability than the auto-regressive objective used in GPT (Devlin et al., 2018). Therefore, we expect conditioned language encoding is more helpful to learn condition-related expressions (especially with the TF-IDF masking strategy which we will introduce) than the two generation tasks that employ the auto-regressive objective.
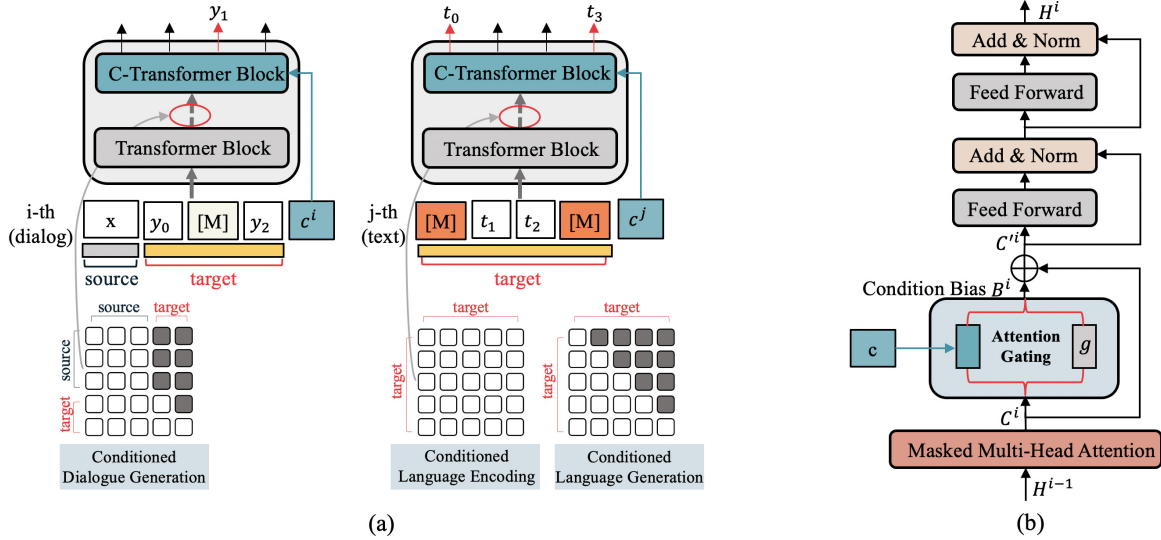
Figure 1: (a) Overview of our multi-task learning approach. Labeled dialogue and text data are mixed, and they are processed using the same pre-trained Transformer with data/task-adaptive input representations, self-attention masks, and random mask strategies. (b) Detailed structures of a condition-aware transformer block, i.e. a C-Transformer Block.

3.2 Condition-aware Transformer Block

In this subsection, we introduce position-wise condition bias that aims to determine how much condition information should be utilized to bias word generation probability at a position. The core component to calculate the bias is a **non-parametric attention-based gating mechanism** as shown in Figure 1 (Right). Other gate mechanisms usually employ parametric linear layers to calculate weights. We assume a non-parametric attention based method could be more training-efficient, which is important since labeled data are usually limited. We will empirically confirm its effectiveness compared to other gating methods.

Specifically, given a training sample (x, c, y) or (c, text), the condition label c is encoded using two sets of parameters: one parametric vector works as the key $\mathbf{k}^c \in \mathbb{R}^{d_h}$ and another one works as the value $\mathbf{v}^c \in \mathbb{R}^{d_h}$. Additionally, there is a general condition label g with a parametric vector $\mathbf{k}^g$ as its key and a zero vector $\mathbf{v}^g$ as its value. The former corresponds to conditioned generation, while the latter to the general dialogue that generates words only based on dialogue history. At each position, the model determines an attention weight to each choice. More attention to c means that the position is more tuned to the condition. More specifically, for each condition-aware transformer block as shown in Figure 1(Right), given $\mathbf{C}^i = [\mathbf{c}_1^i, ..., \mathbf{c}_n^i]$ as queries, the condition biases

$\mathbf{B}^i = [\mathbf{b}_1^i, ..., \mathbf{b}_n^i]$ are calculated by:

$$\mathbf{B}^i = \mathrm{softmax}(\frac{\mathbf{C}^i \mathbf{K}_b^T}{\sqrt{d_k}} + \mathbf{M}_b)\mathbf{V}_b \qquad (3)$$

where $\mathbf{K}_b = [\mathbf{k}^c, \mathbf{k}^g]$ and $\mathbf{V}_b = [\mathbf{v}^c, \mathbf{v}^g]$. The calculation is non-parametric. We use the matrix $\mathbf{M}_b \in \mathbb{R}^{n \times 2}$ to prevent adding condition bias to positions on the source side because the condition only influences the target side (the labeled response or text).

3.3 Objectives

We jointly optimize three tasks: conditioned dialogue generation on labeled dialogue, conditioned language encoding and conditioned language generation on labeled text. As discussed in Section 3.1, conditioned language encoding is expected to be very helpful to learn condition-related expressions.

A specific self-attention mask is required for each task, while the objectives of three tasks are essentially the same – some tokens of the target side (labeled response or text) are randomly masked, and the final hidden vectors H^L corresponding to the masked tokens are fed into an output softmax over the vocabulary to predict the expected tokens. Therefore, we can mix up 2 types of data (3 different tasks) in one training batch, and the loss is averaged in a batch. This thus prevents us from having the catastrophic forgetting problem (Phang et al., 2018). This problem is usually observed

using a sequential fine-tuning process, i.e. first fine-tuning on labeled texts and then on conditioned dialogue data, which will erase the effect of the previous steps of training.

When using labeled dialogue data, we want the model to learn to generate conditioned but more importantly coherent responses. Thus, we uniformly sample the tokens on the target side to mask. Differently, when exploiting labeled text data, we only want the model to generate condition-related expressions. Therefore, we introduce **TF-IDF Based Masking** for the labeled text data to speed up the learning process – we sample tokens to mask according to their TF-IDF values counted on the entire corpus. We will empirically show its effectiveness.

4 Experiments

4.1 Datasets

We use two labeled dialogue datasets, and we created two smaller training sets (500K labeled texts and 250K labeled dialogues), which are summarized in Table 1. We anticipate that when labeled dialogue data are limited, the benefit of leveraging labeled text data will be larger.

Persona Reddit We filtered the Reddit data from 2015 to 2019 that is provided by a third party [3]. Reddit data is a natural source of dialogue with multiple users – a post may have multiple comments by different users. Following Li et al. (2016b), we consider each user as a distinct persona. We extract (post, user, comment) tuples, where "user" is the label of the user who makes the "comment". We further filtered the data based on sentence length and users: sentences with more than 30 words or less than 4 words are removed, and we only keep comments from the 2000 most active users so that we can collect enough data for each user. As a result, each user has 1291 samples (comments) on average. To build the labeled text corpus, we collect extra posts or comments on Reddit from the same user that have no overlap with the dialogue data – these extra texts are intended to reflect the general writing style of the user.

Topic-related Dialogue Dziri et al. (2019) provides a high-quality 3-turns conversational dataset for topic aware response generation [4]. Along with each (history, target) pair, there is a topic label and dozens of topic words that are predicted by LDA

model. The dataset contains 9.2M samples, from which we sample 3M (history, topic, target) tuples as the labeled dialogue corpus. To construct the labeled text data, we sample other 3M tuples and only keep their (topic, target) parts. Note that the topic labels are generated by LDA, and thus it is difficult to obtain the labeled text data from other sources.

Dataset	Persona Reddit		Topic Dialogue	
Source of Labels	Personal ID		LDA	
Number of Labels	2000		190	
Labeled Texts	3M	500K	3M	500K
dialogue Train	3M	250K	3M	250K
dialogue Valid		80K		80K
dialogue Test		10K		10K

Table 1: Key characteristics of the two datasets.

4.2 Baselines

We choose two strong baselines specifically designed for personalized response generation and two others for topic-aware generation. Additionally, we choose some state-of-the-art pre-trained Transformers.

Speaker Model (Li et al., 2016b) a seq2seq model using four LSTM layers. Given a user label, the decoder transforms it into a user embedding and use it to generate a personalized response.

MT-Speaker an approach jointly trains a Speaker Model and a conditioned auto-encoder with shared decoder parameters, which is adapted from a style transfer approach (Luan et al., 2017). This approach also leverages the labeled text data.

TA-Seq2Seq (Xing et al., 2017) and **THRED** (Dziri et al., 2019) these models utilize topic words instead of topic labels predicted by the LDA model. TA-Seq2Seq leverages the topic information by a joint attention mechanism and a biased generation probability. THRED is built based on HRED and incorporates topic words via a hierarchical joint attention mechanism.

C-Trans-ED (Zheng et al., 2019) an encoder-decoder transformer framework initialized with GPT parameters. The decoder dynamically merges features from the dialogue history and the condition. This model is based on the code of ConvAI2 champion (Dinan et al., 2019).

C-Trans-Dec a decoder-only transformer initialized with GPT-2 parameters, adapted from Wolf

et al. (2019). We add a condition embedding to the input representation to enable conditioned generation.

BERT fine-tuning the pre-trained model (Devlin et al., 2018) on the dialogue datasets. The encoder and decoder share the parameters. When encoding, the model uses bi-directional attention. When decoding, it uses left-to-right attention.

4.3 Implementation Details

We implement the speaker model and MT-Speaker model based on OpenNMT [5]. Other models are directly taken from the available open-source code. Hyper-parameters are set following the original papers. Since our baselines utilize GPT or BERT, we use BERT (base, uncased) to initialize our model for fair comparison. It is however possible to build our model upon more powerful pre-trained models such as Roberta(Liu et al., 2019). We do hyper-parameter search based on perplexity on the validation set for: the number of condition-aware transformer blocks in $\{2, 6, 12\}$, the mix-up rate of labeled dialogues and texts in $\{3:1, 1:1\}$, and whether using conditioned language encoding task. We report experimental results with 2, 3:1, and using conditioned language encoding respectively. The warm-up proportion is set to 0.1. 25% tokens of the target side are randomly masked. During decoding the beam size is 10, and we prevent duplicated bigrams. We fine-tune all the parameters end-to-end for four epochs on two P100 GPUs. With in total 6M training samples, each epoch takes twelve hours. The fine-tuning model only has $(2C + 1) \times d_h$ additional parameters, where C is the number of different condition labels. Other details are given in Appendix A.

4.4 Evaluation

Automatic Metrics We choose some widely used metrics in the literature [6]: **BLEU** (Papineni et al., 2002) with n=1,2,3; **ROUGE-L** – longest common subsequence based statistics; **CIDEr** (Vedantam et al., 2015) utilizing TF-IDF weighting for each n-gram; and **Distinct** (Li et al., 2016a) indicating the proportion of unique n-grams (n=1,2) in the entire set of generated responses to evaluate response diversity. Two-sided t-test is used for statistical significance test.

Response Appropriateness Furthermore, we conduct manual evaluation on the best models according to the automatic metrics. We only manually evaluate the model performance on large-scale datasets[7]. We ask human evaluators to rate a response in $\{0, 1, 2\}$. A score of 0 means that the response might have flaw in fluency and logic or be incoherent. Special cases are for example completely coping from the dialogue history as the output, and a bland response such as "I don't know what you mean". A score of 1 represents a coherent but generic response. 2 represents a coherent and informative response. We also do a pair-wise evaluation to compare two models and indicate which one is better. The evaluation is based on 200 random samples. Each generated response is rated by three annotators. The inter-rater annotation agreement in Cohen's kappa (Cohen, 1960) is 0.441 on average, which indicates moderate agreement.

Condition Consistency We observe that automatic metrics fail to evaluate condition consistency since BERT that does not consider conditions outperforms C-Trans-ED and C-Trans-Dec. Thus, we perform manual evaluation on the condition consistency. A generated response is rated in $\{0, 1, 2\}$. The scores 0, 1 and 2 mean respectively that the response is inconsistent to the condition, somehow related, and consistent. However, if the response has flaw in fluency or logic, it will get a score of 0. For Topic Dialogue, it is easy to measure whether a generated response is in the topic. However, for persona consistency, it is difficult for a human evaluator to know the speaking style of each user. Thus, before evaluation we first automatically determine those frequently used words by a user in responses and show them to the annotators to help their evaluations.

4.5 Analysis

Table 2 and 3 gives automatic evaluation results, and Table 4 gives human evaluation results. Appendix B shows some generated responses. The results can be summarized as follow:

BERT vs. Trans-ED & Trans-Dec C-Trans-Dec has a clear advantage over C-Trans-ED in almost all automatic metrics, which can also

---

Model	BLEU-1	BLEU-2	BLEU-3	ROUGE-L	CIDEr	Dist-1	Dist-2	avgLen
Sp-Model	10.539 (**)	3.152 (**)	1.396 (**)	0.116 (**)	0.056 (**)	0.012 (**)	0.044 (**)	12.6
MT-Speaker	10.970 (**)	3.488 (**)	1.540 (**)	0.118 (**)	0.059 (**)	0.009 (**)	0.034 (**)	12.7
C-Trans-ED	13.548 (*)	3.881 (**)	1.529 (**)	0.113 (**)	0.045 (**)	0.005 (**)	0.025 (**)	18.7
C-Trans-Dec	12.964 (**)	4.182 (**)	1.781 (**)	0.117 (**)	0.060 (**)	0.023 (**)	0.097 (**)	16.7
BERT	12.928 (*)	4.405 (/)	1.764 (**)	0.119 (**)	0.062 (**)	0.014 (**)	0.052 (**)	26.1
Ours	**14.052**	**4.891**	2.149	**0.122**	0.070	**0.024**	0.098	23.3
Two-Step FT	13.714 (/)	4.870 (/)	**2.160** (/)	**0.122** (/)	**0.071** (/)	0.023 (/)	0.102 (*)	25.0
w/o ctext	13.015 (*)	4.563 (/)	1.956 (/)	0.113 (**)	0.061 (**)	0.023 (/)	**0.106** (*)	25.7
w/o tfidf	13.581 (*)	4.705 (/)	2.000 (/)	0.118 (**)	0.070 (/)	0.023 (/)	0.095 (*)	24.0
Sp-Model	10.467 (**)	3.039 (**)	1.239 (**)	0.116 (**)	0.049 (**)	0.007 (**)	0.027 (**)	12.3
MT-Speaker	10.286 (**)	2.932 (**)	1.174 (**)	0.114 (**)	0.047 (**)	0.007 (**)	0.030 (**)	12.3
C-Trans-ED	10.968 (**)	3.247 (**)	1.295 (**)	0.106 (**)	0.040 (**)	0.001 (**)	0.006 (**)	14.7
C-Trans-Dec	11.263 (**)	3.390 (**)	1.274 (**)	0.106 (**)	0.043 (**)	0.020 (**)	0.075 (**)	16.2
BERT	12.766 (*)	4.195 (*)	1.805 (*)	0.118 (/)	0.063 (*)	0.022 (/)	0.071 (**)	15.3
Ours	**13.517**	**4.517**	**1.988**	**0.119**	**0.068**	0.021	0.066	16.4
Two-Step FT	10.125 (**)	3.295 (**)	1.388 (**)	0.111 (**)	0.052 (**)	0.015 (**)	0.043 (**)	12.7
w/o ctext	11.776 (**)	3.821 (**)	1.631 (**)	0.115 (**)	0.059 (**)	0.020 (*)	0.062 (**)	14.4
w/o tfidf	13.475 (/)	4.409 (/)	1.853 (/)	0.118 (/)	0.064 (*)	**0.023** (/)	**0.078** (*)	16.7

Table 2: Evaluation results on large-scale (upper half) and small-scale (lower half) Persona Reddit. Two-Step FT means using our model architecture but applying sequential fine-tuning. w/o ctext is without leveraging conditioned text data. w/o tf-idf means without applying TF-IDF based masking. * ($p < 0.05$) or ** ($p < 0.01$) show statistically significant differences with our model by two-sided t-test.

Model	BLEU-1	BLEU-2	BLEU-3	ROUGE-L	CIDEr	Dist-1	Dist-2	avgLen
TA-Seq2Seq	10.197 (**)	3.307 (**)	1.602 (**)	0.121 (**)	0.098 (**)	0.016 (**)	0.051 (**)	9.7
THRED	9.061 (**)	3.035 (**)	1.468 (**)	0.118 (**)	0.098 (**)	0.015 (**)	0.048 (**)	8.8
C-Trans-ED	13.990 (**)	5.359 (**)	2.689 (**)	0.131 (**)	0.147 (**)	0.055 (**)	0.222 (**)	12.5
C-Trans-Dec	14.544 (**)	5.475 (**)	2.669 (**)	0.136 (**)	0.154 (**)	0.046 (**)	0.177 (**)	13.2
BERT	15.287 (/)	6.243 (/)	3.283 (/)	0.141 (/)	0.168 (**)	0.057 (**)	0.227 (**)	12.5
Ours	15.639	**6.484**	**3.455**	0.140	0.185	0.060	0.243	13.0
Two-Step FT	**15.926** (/)	6.431 (/)	3.376 (/)	**0.143** (/)	0.185 (/)	0.059 (*)	0.239 (*)	13.1
w/o ctext	15.491 (*)	6.397 (/)	3.399 (/)	0.142 (/)	**0.190** (*)	**0.063** (*)	**0.262** (**)	12.8
w/o tfidf	15.393 (/)	6.302 (/)	3.351 (/)	0.139 (/)	0.185 (/)	0.059 (**)	0.230 (**)	13.1
C-Trans-ED	13.874 (**)	5.145 (**)	2.503 (*)	0.124 (**)	0.124 (**)	0.039 (**)	0.150 (**)	13.1
C-Trans-Dec	**14.899** (/)	5.648 (/)	2.690 (/)	0.133 (**)	0.150 (/)	0.043 (*)	0.176 (*)	15.2
BERT	14.457 (/)	5.583 (/)	2.802 (/)	0.135 (**)	0.136 (**)	0.037 (**)	0.133 (**)	12.4
Ours	14.587	**5.747**	**2.894**	**0.139**	**0.152**	**0.050**	**0.186**	12.0
Two-Step FT	13.941 (**)	5.463 (/)	2.765 (/)	0.136 (*)	0.140 (**)	0.045 (**)	0.169 (**)	11.7
w/o ctext	13.211 (**)	5.179 (**)	2.655 (/)	0.137 (/)	0.142 (**)	0.046 (**)	0.163 (**)	10.8
w/o tfidf	13.964 (**)	5.485 (**)	2.809 (/)	0.135 (**)	0.145 (*)	0.048 (*)	0.178 (**)	11.8

Table 3: Evaluation results on large-scale and small-scale Topic Dialogue. Topic labels are predicted by LDA.

be observed in their generated responses. Fine-tuning BERT without considering conditions outperforms C-Trans-Dec on most similarity metrics such as BLEU. We explain this by the fact that bi-directional attention could enable a model to better encode dialogue history, and thus to generate responses more similar to the ground truth. The ablation model using w/o ctext is fine-tuning C-BERT (with our condition-aware transformer blocks) on labeled dialogue data. The performance of w/o ctext is similar to C-Trans-Dec's, with a slight advantage in condition consistency and small disadvantage in response appropriateness. These results show that our approach is built upon a strong base model. As mentioned, other pre-trained models can also be used.

With Condition When large Persona Dialogue is available, w/o ctext (i.e. C-BERT) outperforms BERT in almost all automatic metrics. However, we observe that when only small-scale labeled dialogue data are available, all three conditioned models perform worse than BERT. This shows that the model cannot learn the condition-related features well from the limited labeled dialogue data. Thus, it is important to leverage the labeled texts that are easier to collect, and the results on small-scale Persona Reddit show that our multi-task learning approach significantly outperforms BERT on similarity metrics such as BLEU and CIDEr.

For Topic Dialogue, the labels are given by LDA model. LDA is an unsupervised method and the predicted condition labels can be very noisy. Nev-

Model	Persona				Topic			
	Appropriateness		Consistency		Appropriateness		Consistency	
	Score	Pair-wise	Score	Pair-wise	Score	Pair-wise	Score	Pair-wise
C-Trans-Dec	0.96	(28%, 39%)	0.85	(20%, 39%)	0.77	(26%, 34%)	0.71	(21%, 31%)
BERT	0.77	(11%, 40%)	0.78	(22%, 43%)	0.55	(17%, 40%)	0.46	(16%, 40%)
Ours	**1.15**	-	**1.24**	-	**0.83**	-	**0.80**	-
w/o ctext	0.91	(26%, 39%)	0.90	(23%, 38%)	0.73	(27%, 35%)	0.72	(23%, 30%)

Table 4: Human evaluation of generated responses on appropriateness and condition consistency. Pair-wise comparisons show the wining percentages of (baseline, ours).
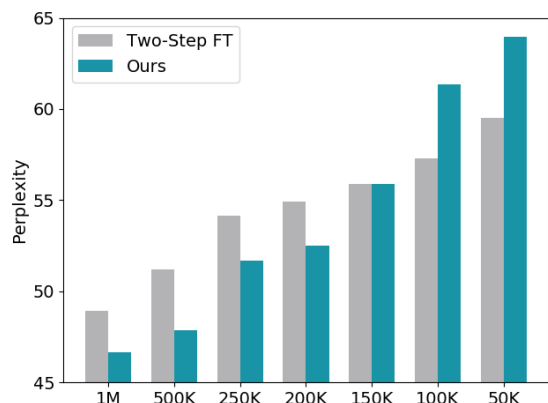

Figure 2: Performance comparison between sequential fine-tuning and our approach given 1M labeled text data and different size of labeled dialogue data.

Model	BLEU-1	BLEU-2	Dist-2
Single Gate	13.880 (*)	4.853 (/)	0.090 (**)
Double Gates	13.988 (*)	4.889 (/)	0.094 (*)
Attn. Routing	**14.052**	**4.891**	**0.098**
Single Gate	11.703 (**)	3.891 (**)	0.090 (**)
Double Gates	11.336 (**)	3.698 (**)	**0.091** (**)
Attn. Gating	**13.517**	**4.517**	0.066

Table 5: Comparison of gating mechanisms on large-scale and small-scale Persona Reddit.

ertheless, similarly, with large data C-BERT outperforms BERT in all metrics, but when only small-scale labeled dialogue data are available, C-BERT performs worse than BERT in terms of BLEU. The result again shows the importance of exploiting labeled texts, and our approach is the best on small-scale Topic Dialogue.

Leveraging Labeled Texts In general, our approach significantly outperforms all baselines and w/o ctext that do not exploit labeled text data, either with large-scale or small-scale data. With small-scale data, our approach outperforms BERT while w/o ctext itself cannot achieve this, which shows that conditioned dialogue generation can be helped by extra labeled text data. On Topic Dialogue, with such noisy labels, our model leveraging the labeled texts still produces the best performance, which confirms the robustness of our multi-task learning approach to work with different types of labels. The human evaluation on appropriateness and condition consistency further confirms the effectiveness of our approach.

Not all methods utilizing extra labeled text can obtain such performance improvement as we did. MT-Speaker that employs an extra auto-encoder does not gain much improvement over Sp-Model. This result shows that using additional model components to leverage labeled texts is inefficient for conditioned dialogue generation. Furthermore, Two-Step FT that first fine-tuning on labeled texts and then on labeled dialogue data does not always produce good performance. It achieves comparable performance to our approach on large-scale datasets, but on small-scale datasets it can even perform worse than w/o ctext (Table 2). This result shows that with small-scale dataset, it is better to avoid sequential fine-tuning because first fine-tuning on labeled texts will erase the effect of the previous step of pre-training. Furthermore, we investigate how the ratio of the size of labeled text data to the size of dialogue data influence model performance. As shown in Figure 2, given 1M labeled text data, when the ratio is less than 6.7, our approach performs better than Two-Step FT. However, when labeled text corpus is much larger than dialogue corpus, sequential fine-tuning is better. We assume that with large labeled text corpus the pre-trained language model can be tuned to conditioned language generation. Besides, the final task in sequential fine-tuning is purely conditioned dialogue generation, which is expected to achieve better performance on dialogue than a multi-task learning approach. However, in real application situations, one cannot always expect that a large labeled text corpus as supplement for the dialogue data is available.

TF-IDF Masking and Attention Gating We assumed that the general language features have already been captured by the pre-trained models. Thus, to better utilize labeled text data, we mask more condition-related words using TF-IDF based masking. Our ablation study confirms that TF-IDF masking brings improvement in almost all automatic metrics although the improvement might not always be statistically significant.

Our attention gating is a non-parametric gating mechanism to fuse the condition into the decoder. We expected it to be efficient, which is particularly important when labeled data are limited. Here, we compare it with two common parametric gating mechanisms: 1) setting a single gate on $\mathbf{C}^i$ to get a weight; 2) setting gates on both $\mathbf{C}^i$ and $\mathbf{v}^c$ to get two weights. Then, we combine the weighted $\mathbf{C}^i$ and $\mathbf{v}^c$ to get $\mathbf{C}'^i$ as in our attention gating. Experimental results in Table 5 confirm that our method is more efficient. When only small-scale labeled data are available, the model with attention gating generates responses that are significantly more similar to the ground-truth.

Conclusion

In this paper, we examined the data scarcity issue of conditioned dialogue generation. Pre-training on unlabeled text or dialogue data is not helpful to conditioned generation. Thus, we exploited labeled text data that are easier to collect than labeled dialogues. We expected these data can contribute to better representations of conditions and better use the conditions in natural language generation, which complement what is lacking in the pre-trained models.

To leverage these two types of data, we proposed a simple and efficient multi-task learning approach. Three tasks are considered: conditioned dialogue generation task on the labeled dialogue data, conditioned language encoding task and conditioned language generation task on the labeled text data. We conducted experiments under persona and topic conditions. Experimental results show that our approach outperforms the state-of-the-art models by leveraging labeled texts, and it also obtains larger improvement in performance comparing to the previous methods leveraging text data.

References

Siqi Bao, Huang He, Fan Wang, and Hua Wu. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating multiple diverse responses with multi-mapping and posterior mapping selection. *arXiv preprint arXiv:1906.01781*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

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. In *Advances in Neural Information Processing Systems*, pages 13042–13054.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating multiple diverse responses for short-text conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6383–6390.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2018. Dialogwae: Multimodal response generation with conditional wasserstein autoencoder. *arXiv preprint arXiv:1805.12352*.

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, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

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 Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2019. Caire: An end-to-end empathetic chatbot. *arXiv preprint arXiv:1907.12108*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 605–614.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2020. The adapter-bot: All-in-one controllable conversational model. *arXiv preprint arXiv:2008.12579*.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

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*.

Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2017. Modeling situations in neural chat bots. In *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2019. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *arXiv preprint arXiv:1911.03768*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. *arXiv preprint arXiv:1903.03094*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yan Zeng and Jian-Yun Nie. 2020. Open-domain dialogue generation based on pre-trained language models. *arXiv preprint arXiv:2010.12780*.

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 (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019. A pre-training based personalized dialogue generation model with persona-sparse data. *arXiv preprint arXiv:1911.04700*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137.

A More Implementation Details

For the small-scale datasets, we trained the models until the performances stop to increase on validation set to prevent over-fitting. For large-scale datasets, we fine-tuned the models for four epochs. In Table 6, the average runtime is tested using a 1080Ti GPU device, and the batch size is set to take all of the GPU memories. TA-Seq2Seq and THRED are implemented in TensorFlow. Other models are implemented in PyTorch. Notice that the runtime will be influenced by code implementation in additional to model structure. When experimenting with the small-scale Persona Reddit dataset, we decrease the number of parameters of Sp-Model and MT-Speaker models to 48M and 52M respectively in order to avoid over-fitting. C-Trans-ED loads the pre-training results of GPT. In the original paper, they pre-trained by themselves using a Chinese corpus, which cannot be used in our experiments.

Model	Parameters	Runtime(min/M)
Sp-Model	80M	25
MT-Speaker	90M	40
TA-Seq2Seq	155M	150
THRED	174M	135
C-Trans-ED	120M	180
C-Trans-Dec	126M	290
BERT	110M	140
Ours	113M	145

Table 6: The number of parameters of each tested approach and the average runtime (minutes) for every million training samples.

Hyper-parameters	Value
C-Tranformer layers	2
mask probability	0.25
max length	80
batch size	160
learning rate	3e-5
warmup proportion	0.1
label smoothing	0
weight decay	0.01
dropout probability	0.1

Table 7: Hyper-parameters for our fine-tuning approach. There are in total 6M data. Thus, we use a large batch size.

Additionally, we explored the idea proposed in Zeng and Nie (2020) to decrease finetune-generation discrepancy introduced by MLM training objective. Nevertheless, the conditioned language encoding task cannot employ this method because it applies bi-directional self-attention. Experimental result on small-scale Persona Reddit shows that eliminating this discrepancy helps to decrease perplexity from 55 to 52.

B Generation Samples

dialogue His.	ca n't wait to see the team at full strength when watt and clowney are back .
Gold Resp.	i 've been waiting a long time to see watt and clowney healthy together
Sp-Model	i do n't think you know what the word means
MT-Speaker	i do n't think the texans will be able to handle the ball in the first half of the season
C-Trans-ED	i do n't know why you 're getting downvoted, i 'm just saying that you 're wrong .
C-Trans-Dec	i 'm not sure if i 'd be surprised if they both get hurt .
BERT	i ca n't wait to see watt and clowney at full strength when watt is back .
Ours	i do n't think it 's going to happen , but i 'd love to see watt and clowney back .
Two-Step FT	i 'd love to see watt and clowney on the same team
w/o ctext	i do n't think it 's going to happen , but i 'd love to see watt and clowney on the same team
w/o tfidf	i do n't think it 's going to happen , but i 'd love to see watt back .
Sp-Model	i do n't think he is going to be a good qb .
MT-Speaker	i 'm not sure if this is a good thing .
C-Trans-ED	i 'm not sure what you 're talking about . i do n't know what to say .
C-Trans-Dec	i 'm not sure i 'd be surprised if they 're both back
BERT	i do n't think they 'll be able to do that .
Ours	i 'd be so happy to see watt and clowney back .
Two-Step FT	i do n't think that 's what i 'm talking about .
w/o ctext	i do n't think they 'll be able to do that .
w/o tfidf	i 'm so excited to see watt on the team .

Table 8: Responses generated by baselines and our model trained on the large-scale and small-scale Persona Reddit.

dialogue His.	what is your favorite food ? [SEP] ice cream .
Gold Resp.	what flavor ?
TA-Seq2Seq	what kind of ice cream ?
THRED	what kind of ice cream ?
C-Trans-ED	ice cream is the best food i have ever eaten
C-Trans-Dec	i 'm not sure if you 're joking but i 'm not sure if you 're serious .
BERT	i do n't like ice cream .
Ours	i do n't like ice cream . it 's coarse and irritating and it gets everywhere .
Two-Step FT	i 've never had ice cream . what 's it like ?
w/o ctext	i 've never had ice cream . what 's it taste like ?
w/o tfidf	what kind of ice cream do you like ?
C-Trans-ED	ice cream is the best food i 've ever seem .
C-Trans-Dec	i 've never had ice cream .
BERT	i 've never had ice cream .
Ours	i do n ' t like ice cream .
Two-Step FT	i like ice cream , but i do n ' t like it .
w/o ctext	i 've never had ice cream , but it 's so good .
w/o tfidf	i ' ve never had ice cream .

Table 9: Responses generated by baselines and our model trained on the large-scale and small-scale Topic Dialogue.