An Investigation of Suitability of Pre-Trained Language Models for Dialogue Generation – Avoiding Discrepancies

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

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

Abstract

Pre-trained language models have been widely used in response generation for open-domain dialogue. These approaches are built within 4 frameworks: Transformer-ED, Transformer-Dec, Transformer-MLM and Transformer-AR. In this study, we experimentally compare them using both large and small-scale data. This reveals that decoder-only architecture is better than stacked encoder-decoder, and both left-to-right and bi-directional attention have their own advantages. We further define two concepts of model discrepancy, which provides a new explanation to the model performance. As discrepancies may hinder performance, we propose two solutions to reduce them, which successfully improve the model performance.

1 Introduction

It has been shown (Wolf et al., 2019) that leveraging a pre-trained Language Model (LM) based on transformer can achieve excellent performance for dialogue generation. Different approaches have been proposed recently, which can be categorized into 4 frameworks: Transformer-ED(Zheng et al., 2019), an encoder-decoder Transformer, Transformer-Dec (Wolf et al., 2019; Lin et al., 2020), Transformer-MLM (Dong et al., 2019) and Transformer-AR (Bao et al., 2019; Shuster et al., 2019). The latter three all utilize a decoder-only architecture. Besides, Trans-Dec uses left-to-right attention for both source and target side, while Trans-MLM and Trans-AR employ bi-directional attention on the source side to encode dialogue history. Due to this difference, Trans-Dec only utilizes left-to-right pre-trained models, e.g. GPT-2 (Radford et al., 2019), while Trans-MLM/AR are based on the pre-trained models applying bi-directional attention (on the source side), e.g. BERT (Devlin et al., 2018). The difference between Trans-MLM and Trans-AR is that Trans-MLM uses masked language modeling while Trans-AR uses auto-regressive objective.

Recent studies have explored pre-training dialogue models using large-scale Reddit/Twitter data (Adiwardana et al., 2020; Roller et al., 2020). It is then straightforward to fine-tune the models for a specific dialogue task. However, in practice, there may not always be enough data for pre-training. In some cases, we still need to exploit a pre-trained LM. For example, some studies do further pre-training for dialogue based on a pre-trained LM (Zhang et al., 2019; Dong et al., 2019; Bao et al., 2019; Shuster et al., 2019), and some studies that do multi-task learning (e.g. on dialogue and question answering) can only fine-tune based on a pre-trained LM (Lin et al., 2020; Zeng and Nie, 2021). Then, a critical question is how to best exploit a pre-trained LM for dialogue generation. On this question, we have contradictory beliefs in the literature: some researchers believe that Trans-Dec is appropriate because it uses a left-to-right language model that corresponds well to the dialogue generation task (Zhang et al., 2019; Lin et al., 2020), while some others (Dong et al., 2019; Bao et al., 2019) show that Trans-MLM/AR fine-tuning BERT can also achieve state-of-the-art performance.

In this study, we aim to address the above question. To do it, we first compare the 4 frameworks with the same setting on 3 datasets, each with large and small scale training data. Our results on large-scale datasets show that Trans-ED that applies the stacked encoder-decoder architecture does not produce competitive results against the others that use a decoder-only architecture. Trans-Dec/AR generate the most appropriate responses. However, according to automatic metrics, Trans-Dec generates most diverse responses while Trans-AR produce responses most similar to the ground-truth. This may be due to the fact that uni-directional attention does not have constraint from the right side context and

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4481–4494
August 1–6, 2021.

thus is more flexible, while bi-directional attention on source side can better model dialogue context. In contrast, the results on small-scale datasets reveal an important aspect, namely, the discrepancies that may occur between the pre-training and the fine-tuning processes. We then try to explain the performances of the 4 frameworks with respect to the discrepancies.

The concept of model discrepancy has been briefly mentioned in Yang et al. (2019) to mean that the model has been trained in a way, but used in a different way. However, the problem has not been investigated in depth. In this work, we go further in this direction and define two discrepancies: **pretrain-finetune discrepancy** which means the differences in architecture and loss function between pre-training and fine-tuning, and **finetune-generation discrepancy** which means that the way the model is used in generation (inference/test) is different from the way it has been trained. For the 4 frameworks, except Trans-Dec, they all have some pretrain-finetune discrepancies. For example, Trans-AR relies on BERT pre-trained using bidirectional attention, but has to limit it to left-to-right attention on the target side during fine-tuning. Only Trans-MLM has finetune-generation discrepancy because of MLM objective: during training, the model input has random masks, while in the generation process, the input does not contain masks.

Discrepancies might affect the model performance since models with such discrepancies cannot best exploit the pre-trained model or employ the fine-tuned model. Our experiments on small-scale datasets show that the performance of Trans-AR that have larger pretrain-finetune discrepancy drops more sharply than Trans-MLM. Trans-Dec/MLM that have small pretrain-finetune discrepancy have clear advantage over other frameworks according to human evaluation. It becomes clear that discrepancies hinder the performance of a dialogue model. To alleviate the problems, we propose 2 approaches to respectively reduce pretrain-finetune and finetune-generation discrepancies of Trans-MLM, aiming at improving its performance. Our experiments show that both methods bring some improvement. In particular, by eliminating finetune-generation discrepancy of Trans-MLM, our approach significantly outperforms previous methods in most automatic metrics, and achieves comparable performance to Trans-Dec in human evaluation that uses much larger dataset for pre-training. These results confirm that discrepancies are indeed an important factor that influences the effectiveness of leveraging a pre-trained LM for a sequence-to-sequence task, and should be alleviated.

The contributions in this work are as follows:

- We compare the four commonly used frameworks that utilize pre-trained language models for open-domain dialogue generation on 3 public datasets each in large and small scale. and we analyze each framework based on the experimental results.
- We introduce the concept of pretrain-finetune discrepancy and finetune-generation discrepancy, and we examine the discrepancies of each framework.
- We propose two methods to reduce discrepancies[1], yielding improved performance. It is the first investigation that shows explicitly the phenomenon of model discrepancy and its impact on performance.

2 Pre-training Based Frameworks

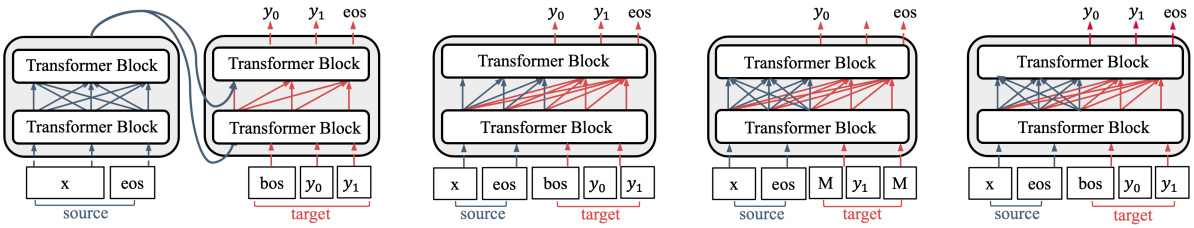
We start with a brief description of the 4 frameworks for dialogue generation based on pre-trained models. More details are provided in Appendix A. We examine the pretrain-finetune discrepancy of each framework. Figure 1 and Table 1 provide an overview.

2.1 Trans-ED

Trans-ED discussed in this paper is an encoder-decoder architecture used by ConvAI2 (Dinan et al., 2019) champion [2]. The decoder of Trans-ED is stacked upon the encoder outputs, while in other decoder-only frameworks, all hidden states of the source side are utilized in the decoding part. The framework shares the encoder and the decoder and initializes the parameters with GPT (Radford et al., 2018). In this case, the pretrain-finetune discrepancy comes from the bi-directional attention in the encoder since GPT is a left-to-right language model. This framework is not commonly used for fine-tuning on a dialogue task. In practice, more efficient variants of Trans-ED are recently used for extremely large-scale dialogue pre-training from

[1]The code is available at: https://github.com/zengyan-97/Transformer-MLM-DiffFree

[2]https://github.com/atselousov/transformer_chatbot



(a) Transformer-ED (b) Transformer-Dec (c) Transformer-MLM (d) Transformer-AR

Figure 1: Architectures of 4 pre-training based Transformers for dialogue generation.

	Trans-ED	**Trans-Dec**	**Trans-MLM**	**Trans-AR**
Pre-trained LM	GPT	GPT-2	BERT	BERT
Architecture	encoder-decoder	decoder-only	decoder-only	decoder-only
Source Side Attn.	bi-directional	left-to-right	bi-directional	bi-directional
Target Side Attn.	left-to-right	left-to-right	left-to-right	left-to-right
Objective	auto-regressive	auto-regressive	MLM	auto-regressive

Table 1: Key characteristics of the 4 pre-training based Transformers. Characteristics in red are inconsistent between pre-training and fine-tuning.

scratch. For example, Adiwardana et al. (2020) utilizes Evolved Transformer to prune redundant connections, and Roller et al. (2020) employs only 2 encoder layers and 24 decoder layers of standard Transformer (Vaswani et al., 2017).

2.2 Trans-Dec

Trans-Dec is a left-to-right decoder-only architecture, and it utilizes GPT-2 (Radford et al., 2019). Thus, there is no pretrain-finetune discrepancy in terms of architecture and loss function. This framework is widely applied for fine-tuning on a dialogue task. However, it encodes dialogue history using only left-to-right attention, which limits the scope of context, resulting in a partial context modeling.

2.3 Trans-MLM and AR

These two frameworks have an identical decoder-only architecture that employs different self-attention masks for the source and target side: they use bi-directional attention on the source side to encode dialogue history and left-to-right attention on the target side. The only difference between them is the objective function: Trans-MLM masks some tokens at the target side and tries to predict them, while Trans-AR uses auto-regressive objective that tries to predict the next tokens successively. BERT is often exploited by the two frameworks, which is a bi-directional architecture using MLM as the pre-training objective. Thus, the pretrain-finetune discrepancy of Trans-MLM/AR comes from the left-to-right attention on the target side. Additionally, Trans-AR applies the auto-regressive objective, which is different from the MLM used in the pre-training.

2.4 Applications of the Frameworks

The four frameworks we described have been widely applied to dialogue generation. For personalized response generation, Wolf et al. (2019) uses Trans-Dec and Zheng et al. (2019) utilizes Trans-ED. Lin et al. (2019) uses Trans-Dec for empathetic response generation. Zeng and Nie (2021) proposes a multi-task learning approach based on Trans-MLM for conditioned dialogue generation. Meanwhile, some studies propose to further pre-train the model using large-scale dialogue data based on a pre-trained language model: Zhang et al. (2019) trains Trans-Dec on 147M Reddit data based on GPT-2, Dong et al. (2019) trains Trans-MLM on natural language understanding and generation datasets based on BERT, Shuster et al. (2019) trains Trans-AR on large-scale Reddit data and then jointly trains on 12 dialogue sub-tasks based on BERT, and Bao et al. (2019) trains a variant of Trans-AR on large-scale Reddit and Twitter data based on BERT. Some recent studies have increased the model size to billions of parameters and utilize even more training data, e.g. Reddit, to train a conversational model from scratch (Adiwardana et al., 2020; Roller et al., 2020; Bao et al., 2020b).

In general, these studies show that all the 4 frameworks can produce good results, and increasing the model size and training data is an effective method to further improve performance. However, behind the success story, the question of suitability

	Twitter	Ubuntu	Reddit
Train Set	2M	1.5M	3M
Valid Set	60K	30K	80K
Test Set	20K	20K	20K

Table 2: Key characteristics of the three public datasets. For each dataset, we also evaluate model performance using **100K** training data and the same test set.

of a framework is masked. To investigate this question, we do not follow the current trend to increase the model size and training data. Instead, we are interested in the behaviors of different frameworks on the same datasets and to understand the reasons.

3 Experiments

3.1 Datasets

We use all the three large-scale unlabeled dialogue datasets in Shuster et al. (2019). Some important characteristics of the datasets are summarized in Table 2. We are interested in the behaviors of the models in two cases: 1) further pre-training on large dialogue data based on a pre-trained LM; and 2) fine-tuning on a small dialogue corpus based on a pre-trained LM. Our large datasets contain a few million samples, and the small datasets consist of 100K samples[3]. Although the datasets are smaller than those used in several previous studies, we believe that a comparison of different models on the same data, and the contrast between large and small datasets, can reveal interesting trends, which we will explain with respect to discrepancies.

Specifically, we choose the following 3 datasets: **Twitter Dialogue Corpus** [4] is collected from Twitter consisting of 2.6M (message, response) pairs. We filtered out samples with history length longer than 72 words (to limit the computation) or shorter than 6 words (not enough information). Samples whose response is longer than 36 words or shorter than 6 words are also removed. As a result, 2M samples are kept. **Reddit Conversational Corpus** [5](Dziri et al., 2019) is a 3-turn conversational dataset collected from 95 selected subreddits. **Ubuntu Dialogue Corpus V2.0** [6] (Lowe et al., 2017) contains two-person conversations extracted from the Ubuntu chat logs of technical support for various Ubuntu-related problems.

3.2 Implementation Details

We use open-source implementations for all four frameworks. Only minor adaptations (e.g. for data loading) have been made. The pre-trained language models used by these frameworks in previous studies have comparable number of parameters ($\sim$ 110M), while the pre-training data are in different scales: Trans-ED $<$ Trans-MLM/AR $<$ Trans-Dec. We assume that the difference is trivial when there are millions of dialogue data. In this study, we use the same data for all the frameworks. More implementation details of each framework and the full comparison among pre-trained LM are given in Appendix C.

We also equip all frameworks with an identical decoding script[7] to avoid extra factor affecting the generation quality, which uses beam search with beam size of 4, prevents duplicated uni-grams, and sets minimum response length that encourages diverse generation as in Roller et al. (2020). The minimum response length is set to make the average length of generated responses match with the average target length of the dataset. Generation results are evaluated after applying an identical word tokenization method. With two P100 GPU devices, the maximum input length is set to 128, and we fine-tune all models for 6 epochs and apply early stopping based on the performance on validation set. Our methods (PF-free and FG-free, which will be described in Section 4.1) do not add parameters or increase runtime in comparison with Trans-MLM.

3.3 Evaluation

Automatic Metrics We compare the similarity between generated responses and ground-truth responses using[8]: **BLEU** (Papineni et al., 2002) evaluating how many n-grams (n=1,2,3) overlapped; **CIDEr** (Vedantam et al., 2015) utilizing TF-IDF weighting for each n-gram. Besides, we evaluate response diversity using **Distinct** (denoted Dist) (Li et al., 2016) that indicates the proportion of unique n-grams (n=1,2) in the entire set of generated responses.

[3]Labeled datasets such as persona (Zhang et al., 2018) and emotion (Rashkin et al., 2019) are usually in similar scale.

[4]https://github.com/Marsan-Ma-zz/chat_corpus

[5]https://github.com/nouhadziri/THRED

[6]https://github.com/rkadlec/ubuntu-ranking-dataset-creator

[7]https://github.com/microsoft/unilm/

[8]We use an open-source evaluation tool: https://github.com/Maluuba/nlg-eval

Model	BLEU-1	BLEU-2	BLEU-3	CIDEr	Dist-1	Dist-2	avgLen
SEQ2SEQ-MMI	10.872 (**)	4.555 (**)	2.259 (/)	**0.119** (/)	0.008 (**)	0.028 (**)	10.6
Trans-ED	15.319 (**)	4.877 (**)	2.037 (**)	0.097 (**)	0.014 (**)	0.063 (**)	19.0
Trans-Dec	14.363 (**)	4.861 (**)	2.120 (*)	0.101 (**)	**0.031** (**)	**0.178** (/)	19.9
Trans-MLM	13.749 (**)	4.253 (**)	1.715 (**)	0.061 (**)	0.018 (**)	0.106 (**)	29.3
Trans-AR	**15.694**	**5.221**	**2.272**	**0.119**	0.029	0.164	18.9
FG-free	15.659 (/)	5.176 (/)	2.200 (/)	0.112 (/)	0.027 (**)	0.147 (*)	18.7
Trans-ED	14.813 (**)	4.249 (**)	1.330 (**)	0.066(**)	0.001 (**)	0.004 (**)	18.4
Trans-Dec	13.805 (**)	4.407 (**)	1.787 (**)	0.092(*)	**0.033** (**)	**0.195** (**)	20.2
Trans-MLM	15.487(**)	4.766(**)	1.814(**)	0.092 (*)	0.016(**)	0.080(**)	19.7
Trans-AR	15.213 (**)	4.700 (**)	1.767 (**)	0.090(**)	0.019(**)	0.091(**)	18.8
PF-free	15.880 (*)	4.970 (*)	1.868 (*)	0.093 (*)	0.022 (**)	0.114 (*)	15.7
FG-free	**16.395**	**5.218**	**2.043**	**0.101**	0.026	0.129	16.2
PF&FG-free	15.714 (*)	4.916 (*)	1.780 (**)	0.093 (*)	0.020 (**)	0.111 (*)	18.4

Table 3: Evaluation results on large-scale (upper half) and small-scale (lower half) Twitter dataset. PF-free denotes the method with reduced pretrain-finetune discrepancy of Trans-MLM. FG-free denotes the method that eliminates finetune-generation discrepancy of Trans-MLM. Two-sided t-test compares each method with the one without () sign, which is usually the best performer. Scores are denoted with * ($p < 0.05$) or ** ($p < 0.01$) for statistically significant differences, and / for insignificant differences.

Model	BLEU-1	BLEU-2	BLEU-3	CIDEr	Dist-1	Dist-2	avgLen
SEQ2SEQ-MMI	12.056(**)	5.512(**)	2.841(**)	0.142(**)	0.005(**)	0.024(**)	9.8
HRED-MMI	13.518(**)	4.564(**)	1.947(**)	0.060(**)	0.001(**)	0.003(**)	13.6
Trans-ED	19.295(/)	6.712(**)	2.986(*)	0.125(**)	0.010(**)	0.069(**)	16.8
Trans-Dec	18.974(*)	6.911(/)	3.022(*)	0.130(*)	**0.018**(**)	**0.134**(**)	18.0
Trans-MLM	17.574(**)	5.884(**)	2.552(**)	0.096(**)	0.012(**)	0.097(**)	25.5
Trans-AR	**20.103**	**7.270**	**3.339**	**0.143**	0.017	0.127	16.8
FG-free	19.774 (/)	7.045 (/)	3.213 (/)	0.139 (/)	0.016 (*)	0.115 (/)	17.7
Trans-ED	14.195(**)	4.533(**)	1.756(**)	0.074(**)	0.003(**)	0.012(**)	16.3
Trans-Dec	17.944(**)	6.360(*)	2.727(*)	0.121(/)	**0.018**(**)	**0.143**(**)	18.3
Trans-MLM	18.338(*)	6.018(**)	2.480(**)	0.108(**)	0.011(**)	0.066(**)	17.0
Trans-AR	19.005 (*)	6.431 (/)	2.733 (*)	0.114(*)	0.012(**)	0.078(**)	17.4
PF-free	**19.116** (*)	6.356 (*)	2.684 (*)	0.118 (/)	0.012 (**)	0.086 (*)	16.7
FG-free	18.884	**6.530**	**2.869**	**0.125**	0.014	0.095	17.3
PF&FG-free	19.024 (*)	6.448 (/)	2.740 (*)	0.118 (/)	0.012 (**)	0.087 (*)	17.1

Table 4: Evaluation results on large-scale (upper half) and small-scale (lower half) Ubuntu dataset.

Human Evaluation Furthermore, we ask human evaluators to rate a response in $\{0, 1, 2\}$. 2 represents a coherent and informative response. Details are given in Appendix D. We also do a pair-wise evaluation to compare two models and indicate which one is better. To reduce time cost, we only perform human evaluations on Twitter and Reddit datasets that are closer to daily dialogue. However, during evaluation, we observe that $\sim 65\%$ Reddit data are professional discussions that are difficult to understand. The percentage is $\sim 30\%$ for Twitter data. These test samples are discarded, and at the end the test set for each dataset consists of 200 random samples. The inter-rater annotation agreement in Cohen's kappa (Cohen, 1960) is 0.44 and 0.42 for Twitter and Reddit, which indicates moderate agreement.

In addition to the 4 frameworks, we also include two general RNN-based baseline frameworks – SEQ2SEQ-MMI (Li et al., 2016) and HRED-MMI (Serban et al., 2016) to show how pre-trained models perform against them.

3.4 Architecture Analysis

We first examine architecture appropriateness on the large-scale data setting, since when data are limited pretrain-finetune discrepancy and the size of pre-training data may strongly influence the results. Appendix E shows some generation samples. Our global observation is that Trans-Dec and Trans-AR are the best choice for large-scale data setting, e.g. further dialogue pre-training based on a pre-trained LM.

Left-to-Right Only vs. Bi-Direction on the Source Human evaluation results in response appropriateness (Table 6 and 7) show that Trans-Dec and Trans-AR generate most appropriate responses. According to automatic metrics, Trans-AR applying bi-directional attention on the source side obtains the highest BLEU and CIDEr scores on all

Model	BLEU-1	BLEU-2	BLEU-3	CIDEr	Dist-1	Dist-2	avgLen
SEQ2SEQ-MMI	15.550(**)	6.814(**)	3.321(**)	0.168(**)	0.011(**)	0.036(**)	11.2
HRED-MMI	13.278(**)	3.845(**)	1.398(**)	0.047(**)	0.001(**)	0.003(**)	13.8
Trans-ED	17.946(/)	6.626(**)	3.213(**)	0.165(**)	0.039(**)	0.203(**)	18.8
Trans-Dec	17.581(**)	6.790(*)	3.372(*)	0.180(**)	0.043(/)	**0.248**(**)	18.2
Trans-MLM	18.672(**)	7.115(**)	3.484(/)	0.177(**)	0.041(**)	0.215(**)	16.8
Trans-AR	**18.849**	**7.245**	**3.662**	**0.192**	**0.044**	0.235	16.8
FG-free	18.741 (/)	7.134 (**)	3.504 (*)	0.184 (*)	0.042 (**)	0.225 (**)	17.0
Trans-ED	17.337(**)	5.366(**)	1.967(**)	0.073(**)	0.001(**)	0.003(**)	17.1
Trans-Dec	17.460(**)	6.586(**)	3.161(*)	0.172(/)	**0.045**(/)	**0.254**(**)	17.7
Trans-MLM	19.193 (/)	6.877 (/)	3.175(*)	0.152(**)	0.029(**)	0.128(**)	15.0
Trans-AR	18.749(/)	6.746(/)	3.119(*)	0.153(**)	0.031(**)	0.141(**)	16.2
PF-free	18.466 (/)	6.688 (*)	3.075 (*)	0.169 (*)	0.038 (/)	0.180 (*)	14.1
FG-free	18.610	**6.937**	**3.302**	**0.175**	0.040	0.191	14.1
PF&FG-free	**19.302** (*)	6.923 (/)	3.073 (*)	0.159 (**)	0.034 (*)	0.164 (**)	15.3

Table 5: Evaluation results on large-scale (upper half) and small-scale (lower half) Reddit dataset.

Model	Score (M)	Score (K)
SEQ2SEQ-MMI	0.39	-
Trans-ED	0.53	0.11
Trans-Dec	**1.02**	0.77
Trans-MLM	0.88	0.58
Trans-AR	0.99	0.47
PF-free	-	0.52
FG-free	0.91	**0.78**
PF&FG-free	-	0.72
	Trans-Dec (M)	FG-free (K)
SEQ2SEQ-MMI	(11%, 48%)	-
Trans-ED	(14%, 46%)	(4%, 47%)
Trans-Dec	/	(24%, 29%)
Trans-MLM	(24%, 34%)	(18%, 31%)
Trans-AR	(27%, 32%)	(17%, 34%)
PF-free	-	(18%, 38%)
FG-free	(28%, 32%)	/
PF&FG-free	-	(23%, 29%)

Table 6: Human evaluation including pair-wise evaluation (lower half) for generated response quality for million-scale (M) Twitter dataset and its 100K training subset (K). Pair-wise comparisons show the wining percentages of the two parties.

Model	Score (M)	Score (K)
SEQ2SEQ-MMI	0.12	-
Trans-ED	0.33	0.10
Trans-Dec	0.58	**0.43**
Trans-MLM	0.48	0.38
Trans-AR	0.64	0.31
PF-free	-	0.28
FG-free	**0.68**	0.40
PF&FG-free	-	0.33
	FG-free (M)	Trans-Dec (K)
SEQ2SEQ-MMI	(5%, 40%)	-
Trans-ED	(11%, 33%)	(2%, 28%)
Trans-Dec	(25%, 32%)	/
Trans-MLM	(18%, 29%)	(15%, 19%)
Trans-AR	(18%, 23%)	(15%, 23%)
PF-free	-	(15%, 24%)
FG-free	/	(23%, 24%)
PF&FG-free	-	(16%, 24%)

Table 7: Human evaluation on Reddit dataset.

three million-scale datasets. We believe that bidirectional attention helps the model to better encode the dialogue history. In contrast, Trans-Dec is able to generate the most diverse responses. We attribute it to the left-to-right attention that introduces less constraints than bidirectional attention, thus has a higher flexibility for generation.

Trans-MLM vs. AR With large data, Trans-AR substantially outperforms Trans-MLM in terms of both automatic and human evaluation. When eliminating the finetune-generation discrepancy of Trans-MLM, i.e. FG-free (we will introduce in Section 4.2), the performance is improved while still having a small gap especially in automatic metrics to Trans-AR. This may be because MLM objective only masks a certain percentage of tokens (40%) while AR objective predicts all tokens on the target side for training. Thus, the AR objective is more training-efficient. Similar observation about the efficiency of MLM has been reported in Clark et al. (2020). However, when training data are limited, we will show that it is better to use MLM objective which has smaller pretrain-finetune discrepancy.

Trans-ED vs. Decoder-Only With large dialogue data, we assume the size of pre-training data and pretrain-finetune discrepancy only have small influence on performance. However, even comparing with Trans-MLM(FG-free)/AR, Trans-ED generates much less diverse or appropriate responses. We also observe lower speed for convergence when training the model [9]. We believe that the result is more or less due to the main difference in architecture: an explicit encoder in Trans-ED might be

[9]Similar observation has been reported in: https://github.com/atselousov/transformer_chatbot/issues/15

redundant (Liu et al., 2018).

3.5 Discrepancy Impact

In section 2, we have discussed the pretrain-finetune discrepancy of each framework. When a large training dataset is available, the impact of pretrain-finetune discrepancy is less severe since the model can be gradually adapted to the given task. However, if the training data are limited, the discrepancy problems may surface. Evaluation results, especially in human evaluation, show that the performance is more reduced with small data if the framework has larger discrepancy. For example, by comparing Trans-MLM (FG-free) and Trans-AR, the latter having additional pretrain-finetune discrepancy due to its auto-regressive objective, we see that the performance of Trans-AR drops more when trained on a small dataset. Trans-MLM (FG-free) and Trans-Dec that have small pretrain-finetune discrepancy have clear advantage over other frameworks according to human evaluation.

These results suggest that with a small dataset one should reduce pretrain-finetune discrepancy to best exploit pre-trained LM. In the next section, we propose 2 methods to reduce pretrain-finetune discrepancy and finetune-generation discrepancy of Trans-MLM.

4 Discrepancy-Free Trans-MLM

4.1 Pretrain-Finetune Discrepancy

The discrepancy of Trans-MLM comes from the left-to-right attention on the target side that has not been pre-trained in BERT. Therefore, this discrepancy cannot be eliminated during fine-tuning for a generation task. However, we can alleviate the discrepancy by using bi-directional attention also on the target side. Specifically, at inference time, to generate a new token denoted as g_t, [MASK] is fed into t-th position, denoted as g_t-M. Previously generated tokens $g_{<t}$ could be viewed as a special type of dialogue history, and thus we can apply bi-directional attention on it.

However, in this case, the corresponding training process will have efficiency problems – only one token can be masked in each training sample; otherwise, there will be conflict for the self-attention mask (Appendix B). This would lead to much lower training efficiency: the loss on validation set only decreases slightly to 5.39 from 6.27 after four epochs, while Trans-MLM masking 40% of the target tokens can reduce it to 4.35. To avoid

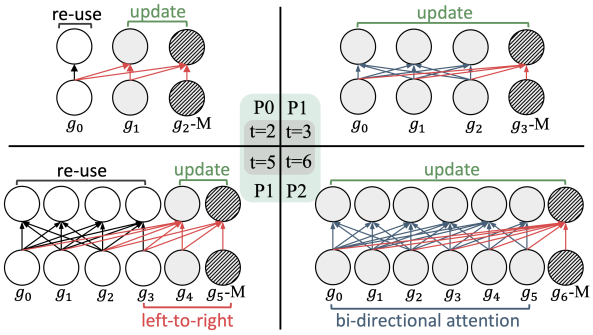


Figure 2: The generation process of PF-free at 4 different time steps. Bi-attention interval is 3 in the graph.

this situation, we cannot always update previous hidden states using bi-directional attention in generation. Therefore, we explore to set a time-step interval for bi-directional attention on the target side – within the interval we apply left-to-right attention and at the end of an interval we apply bi-directional attention. The corresponding training method allows us to mask multiple target tokens at the same time to guarantee training efficiency.

Figure 2 illustrates the generation process of our method with interval of 3. Before time step 3, left-to-right attention is used (e.g. t=2). At time step 3, bidirectional attention is allowed. Then left-to-right attention is used (e.g. t=5) before the end of next interval cycle (t=6). Accordingly, the training process is: given a target response, we first randomly select among all (3 in the figure because t=3 and t=5 are the same pattern) possible attention patterns (e.g. the case of t=3 or t=5 in Figure 2, where we apply bi-directional attention only on $y_{0,1,2}$); then in the part of left-to-right attention, we randomly mask several tokens. We can mask multiple tokens because this part applies left-to-right attention and the masks at other positions will not influence the prediction on a given mask. We call this method **PF-free**, which means that the pretrain-finetune discrepancy is reduced.

4.2 Finetune-Generation Discrepancy

A model having finetune-generation discrepancy means the way that it is used in generation (inference/test) is different from the way it has been trained. Only Trans-MLM has finetune-generation discrepancy because of its MLM objective as shown in Figure 3: during training, there is a masked token, y_1-M, before y_2-M, while in inference there is not a masked token before when generating the token for g_2-M.

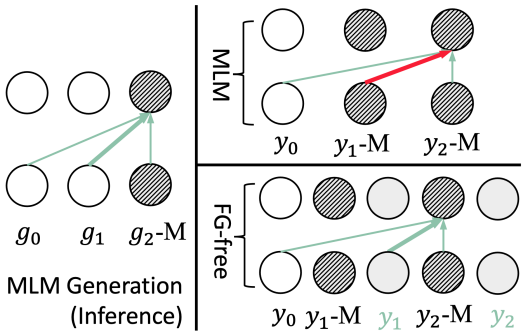


Figure 3: The training process of vanilla Trans-MLM and FG-free. We only plot the attention connection at the second position.

To deal with the problem, we propose that at training time, rather than replacing the tokens with [MASK] as in vanilla MLM, we keep all original input tokens unchanged and prepend [MASK] tokens in the input sequence as illustrated. The prepended [MASK] token uses the same position embedding of the corresponding token. Then, every position after y_1-M attends to y_1 instead of the [MASK] token, and thus the finetune-generation discrepancy of MLM is eliminated. We call the modified model **FG-free**. A similar method has also been explored in (Bao et al., 2020a), where they introduced an extra pseudo mask in addition to [MASK] and prepend it before the original token in order to handle factorization steps of their partially auto-regressive language model.

4.3 Experimental Results

The results with PF-free, FG-free and PF&FG-free models on small-scale datasets are reported in previous tables together with other models. We can see that each of the proposed methods brings some improvement. PF-free improves most automatic metrics over Trans-MLM, but the response appropriateness in human evaluation is not improved. We observe that PF-free could generate some responses that lack fluency, which also influences PF&FG-free (Appendix E). In general, our exploration shows that the left-to-right attention on the target side is necessary for a generative task.

We examine our FG-free method on both large and small-scale data. It always brings statistically significant improvement over Trans-MLM in all automatic metrics, and generates more appropriate responses. On small-scale datasets, it outperforms all other frameworks in similarity metrics and achieve comparable performance in response appropriateness to Trans-Dec that has leveraged much more pre-training data.

This set of experimental results confirm the usefulness of reducing discrepancies in the model. This demonstrates that model discrepancies are indeed important problems we need to address when a pre-trained LM is used for dialogue generation, and the problems have been under-explored.

Conclusion

In this paper, we examined the 4 frameworks for open-domain dialogue based on pre-trained models. We compared their performances on several datasets with the same setting. The comparison revealed that Trans-Dec and Trans-AR are both good choices when large-scale data are available, e.g. further dialogue pre-training. When data are limited, e.g. fine-tuning on small dialogue tasks, Trans-Dec is the most appropriate.

Furthermore, we defined the concept of pretrain-finetune and finetune-generation discrepancy, and examined the 4 frameworks with respect to these concepts. We have shown that the performances of the 4 frameworks can be largely explained by their respective discrepancies, which hinder their performances. This becomes more clear when the dataset is small.

To further show that reducing the discrepancies can improve the performance, we designed PF-free and FG-free correction methods to reduce the discrepancies on Trans-MLM, and tested the corrected Trans-MLM models on the datasets. Our results confirmed that once discrepancies are eliminated, Trans-MLM can produce better results.

This study is the first investigation on the widely used 4 frameworks based on pre-trained LM in terms of architectural appropriateness and discrepancies. We believe that this question is important to understand how a pre-trained model can be used in dialogue generation. It deserves more investigations in the future.

Acknowledgments

This research work is partly supported by an NSERC discovery grant.

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*.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, et al. 2020a. Unilmv2: Pseudo-masked language models for unified language model pre-training. *arXiv preprint arXiv:2002.12804*.

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. 2020b. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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*.

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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. *arXiv preprint arXiv:2004.03829*.

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*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

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.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth 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*.

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.

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*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Yan Zeng and Jian-Yun Nie. 2021. A simple and efficient multi-task learning approach for conditioned dialogue generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4927–4939.

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*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

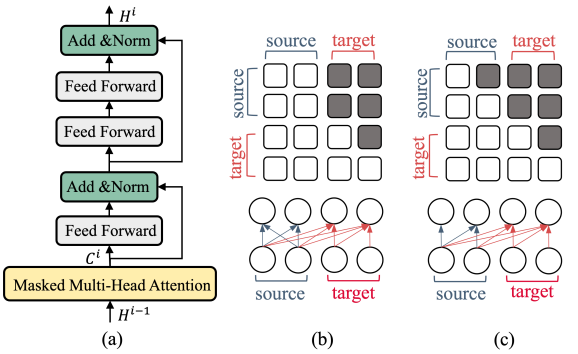


Figure 4: i-th Transformer Block and two $\mathbf{M}$ settings represented in two ways. Shaded areas are blocked.

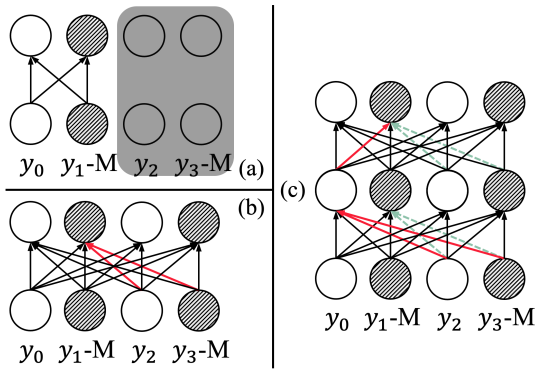


Figure 5: Self-attention mask, $\mathbf{M}$, conflicts – (a) if predicting y_1, y_2 and y_3-M are "future" and forbidden to be accessed by y_1-M; (b) if predicting y_3, y_1-M accesses to y_2 and y_3-M, which causes conflicts to $\mathbf{M}^{(a)}$; (c) if forbidding y_1-M to access to y_2 and y_3-M in $\mathbf{M}^{(b)}$, there will still be (indirect) information leak as indicated in red arrows (y_2 and y_3-M $\rightarrow y_0 \rightarrow y_1$-M). Masking two positions thus causes conflicts. Our PF-free method aims to overcome this problem.

A Multi-Layer Transformer

In this section, we provide some background knowledge on Transformer. The four frameworks we discussed all consist of 12 Transformer blocks. Figure 4 (a) shows a general architecture of a Transformer layer, where the most important component is the masked multi-head self-attention. The setting of attention masks is the largest difference between Trans-Dec and Trans-AR, and it is also the most critical part to implement our PF-free and FG-free methods.

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. Then, $\mathbf{H}^0$ is encoded into hidden representations of i-th layer $\mathbf{H}^i = [\mathbf{h}_1^i, ..., \mathbf{h}_n^i]$ by: $\mathbf{H}^i = \text{Trans}^i(\mathbf{H}^{i-1}), \quad i \in [1, L]$, where Trans^i denotes the i-th Transformer Block as shown in Figure 4 (a). The core component of a transformer block is the masked multi-head attention, whose outputs are $\mathbf{C}^i = [\mathbf{c}_1^i, ..., \mathbf{c}_n^i]$ that are computed via $\mathbf{C}^i = \text{Concat}(\mathbf{head}_1, ..., \mathbf{head}_h)$, with

$$\mathbf{head}_j = \text{softmax}(\frac{\mathbf{Q}_j \mathbf{K}_j^T}{\sqrt{d_k}} + \mathbf{M})\mathbf{V}_j \quad (1)$$

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 $\mathbf{W}_j^Q, \mathbf{W}_j^K, \mathbf{W}_j^V \in \mathbb{R}^{d_h \times d_k}$ respectively. $\mathbf{M} \in \mathbb{R}^{n \times n}$ is the **self-attention mask matrix** that determines whether a position can attend to other positions. $\mathbf{M}_{ij} \in \{0, -\infty\}$. In particular, $\mathbf{M}_{ij} = 0$ allows the i-th position to attend to j-th position and $\mathbf{M}_{ij} = -\infty$ prevents from it. Figure 4 (b&c) shows two $\mathbf{M}$ settings that are applied by Trans-MLM/AR and Trans-Dec respectively.

B Illustration of Attention Conflict

If applying bi-directional attention at each generation step, only one token at the target side could be masked for each training sample; otherwise there will be attention conflicts, i.e. different self-attention mask matrices are required for different masked tokens, while only one mask matrix can be provided per training sample. In Figure 5, we provide an illustration of the mask conflict problem. We assume y_1 and y_3 are masked and need to be predicted at the same time. We see in the figure that two different masks are required for predicting y_1 and y_3, which cannot be done in a single training step, making it impossible to mask more than one token in each step.

C Implementation Details

For the 4 frameworks, we used open-source implementations. Only some minor adaptations to our data and task are made (e.g. re-wrote the data loader to load our experimental datasets, and modified the training objective by keeping only the response reconstruction loss). For response generation, we equipped all frameworks with an identical decoding script [10]. We did not modify other parts, and used the default settings for hyper-parameters, e.g. optimizer and learning rate. Some generation examples are given in Appendix E. Although some models (e.g. Trans-ED) produced poor per-

[10] https://github.com/microsoft/unilm/

Model	Pre-trained LM	Data
Trans-ED	GPT (Radford et al., 2018)	BooksCorpus
Trans-Dec	GPT-2 small (Radford et al., 2019)	WebText
Trans-MLM/AR	BERT base (Devlin et al., 2018)	BooksCorpus, English Wikipedia

Table 8: The text data used for language model pre-training.

formance on small datasets, all model can generate some coherent and fluent responses with large scale training data, which is consistent with the performances reported in previous papers.

Language Models The pre-trained language models used by these frameworks have comparable number of parameters as listed in Table 9, while the pre-training data are in different scales as described in Table 8. BooksCorpus (Zhu et al., 2015) (800M words) contains over 7,000 unique unpublished books from a variety of genres. English Wikipedia (2,500M words) consists of the text passages of Wikipedia extracted by Devlin et al. (2018). WebText crawled by Radford et al. (2019) contains 8M diverse documents for a total of 40 GB of text.

Trans-ED We use the implementation of ConvAI2 champion [11]. The model was for persona-conditioned dialogue generation. The framework is based on GPT architecture and uses GPT for parameter initialization. However they only provide a model checkpoint that has been fine-tuned on large-scale dialogue data including Reddit. To examine the ability of utilizing pre-trained LM, we did not use this checkpoint but initialize the model with GPT parameters [12]. We also did not apply post-processing to the generation results (to be consistent with other experiments).

Trans-Dec We use the released code of Wolf et al. (2019)[13] that uses GPT-2 small by default. The model was for persona-conditioned dialogue generation.

Trans-MLM/AR These two models are implemented based on Dong et al. (2019) [14] that applies multi-task learning on language understanding and generation tasks. We use BERT (base, uncased) for parameter initialization, and fine-tune it on dialogue datasets. PF-free and FG-free are also implemented based on the code. We set the bi-directional attention interval of PF-free to 5. Since the average length of ground-truth responses in the datasets is ~ 15, This setting is generally appropriate.

[11] https://github.com/atselousov/transformer_chatbot

[12] https://github.com/openai/finetune-transformer-lm/tree/master/model

[13] https://github.com/huggingface/pytorch-openai-transformer-lm

[14] https://github.com/microsoft/unilm/tree/master/unilm-v1

Model	Params	Runtime
SEQ2SEQ-MMI	66M	50
HRED-MMI	58M	25
Trans-ED	117M	180
Trans-Dec	117M	290
Trans-MLM	110M	140
Trans-AR	110M	140
PF&FG-free	110M	140

Table 9: The number of parameters of each tested approach and the average runtime (minutes) for every million training samples. The runtime is tested using a 1080Ti GPU device, and the batch size is set to take all of the GPU memories. Notice that the runtime will be influenced by code implementation in addition to model structure.

D Human Evaluation Details

Some existing studies considered response fluency, coherence, and informativeness. We make the manual evaluation simpler and ask the human evaluators to rate a response in $\{0, 1, 2\}$. A score of 0 represents an unacceptable response, which 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 do not know what you are talking about , but it 's a good point .". A score of 1 represents an acceptable response, but it is generic or not perfectly coherent to the dialogue history. 2 represents a coherent and informative response.

E Generation Examples

Dialogue History	one week before election day , early voting is nearly twice as high as <num>
Gold Response	i hope it 's <person> out in full force .
SEQ2SEQ-MMI	i am not voting for a third party candidate . (1)
Trans-ED	i am not sure what you are talking about , but it 's a good thing that <person> is going to win . (1)
Trans-Dec	that 's not true . early voting in nyc is times higher than the national average (2)
Trans-MLM	it 's not even close to <num> % of the vote . i am sure you are right , but there is a huge difference between early voting and <person> voter suppression in ca (1)
Trans-AR	it 's not that high , but i am sure there will be a lot of voter fraud . (2)
FG-free	it 's not even close to <num> yet , but i am sure the polls will show up . (2)
Trans-ED (K)	i am not sure what you are talking about . it 's the only one who is going to do with this (0)
Trans-Dec (K)	it 's not that early voting is a bad thing , but it shouldnot be the case . (0)
Trans-MLM (K)	it 's a little too late . i am not sure if that is the case , but we will see what happens next (2)
Trans-AR (K)	it 's not even close to <num> % of voter turnout . i am sure you are right about that (0)
PF-free (K)	<num> % of people voted for <person> , so i am going to vote (2)
FG-free (K)	it 's not even close to <num> % of voter turnout , but i am sure there are more . (2)
PF&FG-free (K)	it 's not even <num> % of the vote . i am going to be a voter (2)

Table 10: Responses generated by baselines and our methods trained on the **Twitter** dataset(million-scale and 100K). Human evaluation scores are given at the end of each generated reply.

Dialogue History	i think about this man every day
Gold Response	it is so hypnotic . this man was found dead post-election
SEQ2SEQ-MMI	i do not know what you are talking about . (0)
Trans-ED	i think he 's the only one who has a lot of work to do with it . (1)
Trans-Dec	i think about this man every day too . he 's my favorite person in the world (2)
Trans-MLM	he 's such a good guy . i love him so much, but it doesnot make sense to me that this is the only way we can be together in our lives (2)
Trans-AR	i think about this man every day too . he 's so good at it ! (2)
FG-free	he 's such a good guy . i am so happy for him and his family (2)
Trans-ED (K)	i am not sure what you are talking about , but it 's a good idea . (0)
Trans-Dec (K)	i think about this man every day . he 's my hero and i love him (2)
Trans-MLM (K)	i think he 's a great actor , but no one cares about him . lol (2)
Trans-AR (K)	i think he 's a good guy , but not sure what you are talking about . (0)
PF-free (K)	i think he 'snitch . it makes me happy to watch him (0)
FG-free (K)	he 's a great guy . i am so happy to have him back (2)
PF&FG-free (K)	i think he is little silly , but you are a fool . (2)

Table 11: Responses generated by baselines and our methods trained on the **Twitter** dataset(million-scale and 100K). Human evaluation scores are given at the end of each generated reply.

Dialogue History	how old are you guys ? you did not have mechanical pencils ? [SEP] <url> young are you ? i was in school all the way up to a few years ago and most of us still used non - mechanical ones .
Gold Response	i started using mechanical pencil in elementary school about 15 years ago .
SEQ2SEQ-MMI	i do not know what you are talking about . (0)
Trans-ED	how old are you ? i was in school all the way up to a few years ago . (0)
Trans-Dec	i was in school all the way up to a few years ago and most of us still used non - mechanical ones . (0)
Trans-MLM	i was in school for a few years , and most of us still use mechanical pencils . it 's not the same age as you (2)
Trans-AR	i was in high school when we used to use mechanical pencils . (2)
FG-free	how old are you ? i was in high school and still use mechanical pencils . (2)
Trans-ED (K)	i do not know what you are talking about , but it 's just a joke . (0)
Trans-Dec (K)	i was in school all the way up to a few years ago and most of us still used non - mechanical ones . (0)
Trans-MLM (K)	i have been in school for a few years , but not really . (0)
Trans-AR (K)	how old are you ? i have never been in school for a few years . (0)
PF-free(K)	how old are you ? i am not really a mechanical engineer . ! (1)
FG-free (K)	how old are you ? i have never had a mechanical pencil . (2)
PF&FG-free (K)	how old are you ? i have never had a lot of time in school (2)

Table 12: Responses generated by baselines and our methods trained on the **Reddit** dataset(million-scale and 100K). Human evaluation scores are given at the end of each generated reply.

Dialogue History	i have read all of the books , which character do you not like ? [SEP] it 's gon na be the dornish and greyjoys . and probably brienne . you will see .
Gold Response	nah , they have only just started a game of thrones so it must be one of the starks , dan , or tyrion .
SEQ2SEQ-MMI	i do not know why you are being downvoted . (0)
Trans-ED	i have read all of the books , and it 's gon na be dornish . (2)
Trans-Dec	i have not read the books yet , but i will definitely check them out . (2)
Trans-MLM	i have not read the dornish and greyjoys , but you will see them all as well . (1)
Trans-AR	brienne is not a fictional character , she 's an actual person . (2)
FG-free	i have read all of the dornish and greyjoys . (1)
Trans-ED (K)	i am not sure what you are talking about , but it 's just a joke . (0)
Trans-Dec (K)	i do n't know , i have read all of the dornish books . (1)
Trans-MLM (K)	i do not think it 's a good idea to say that . (0)
Trans-AR (K)	i do not think it 's gon na be the best . (0)
PF-free (K)	i am not sure what you are talking about here 's a link . (0)
FG-free (K)	i do not think it 's gon na be . (1)
PF&FG-free (K)	i do not know , but it 's a lot of people ? (1)

Table 13: Responses generated by baselines and our methods trained on the **Reddit** dataset(million-scale and 100K). Human evaluation scores are given at the end of each generated reply.

Dialogue History	is that easy to do ? [SEP] usually the ubuntu installer will set it up perfectly for you
Gold Response	and then i would select at start-up ?
SEQ2SEQ-MMI	how do i install it ?
HRED-MMI	i am not sure what you mean by that
Trans-ED	is there a way to do it without installing the ubuntu installer ? i am not sure .
Trans-Dec	ok , i will try that . thanks for the help ! :-) it 's working now
Trans-MLM	is there a way to do it with the live cd ? i am not sure if that 's what you are looking for , but how does one install ubuntu on
Trans-AR	ok , i will give it a try . thank you very much for your help !
FG-free	is there a way to do it from the command line ? i am not sure what you mean
Trans-ED (K)	i am not sure what you mean by that , but it 's just a little bit more specific
Trans-Dec (K)	ok , i will try that . thanks for your help ! :-) <url>
Trans-MLM (K)	i am not sure what you mean by that , but it 's a good idea .
Trans-AR (K)	i am not sure what you are talking about , but it 's possible that the installer will be able to boot from cd
PF-free (K)	i am not sure what to do ? it ' s just a file in the repos
FG-free (K)	i do not think it ' s a good idea to install ubuntu on my laptop
PF&FG-free (K)	i am not sure what you mean by setting it up perfectly . is there a way to do that ?

Table 14: Responses generated by baselines and our methods trained on the **Ubuntu** dataset(million-scale and 100K).

Dialogue History	usually if on the same machine , <url> [SEP] : i am getting error . phpmyadmin not in this server
Gold Response	you installed it from the software center , or with aptitude or apt-get , or some other way ?
SEQ2SEQ-MMI	what is the error you are getting ?
HRED-MMI	i have no idea what you are talking about
Trans-ED	i am not sure what you are trying to do , but it 's a good idea .
Trans-Dec	did you install phpmyadmin from the repos ? if so , it should be in <path>
Trans-MLM	i am not sure what you are trying to do , but it should be in <path> . try sudo dpkg - reconfigure phpmyadmin
Trans-AR	i am not sure what you are trying to do , but it 's possible .
FG-free	i am not sure what you are trying to do , but it ' s usually in <path>
Trans-ED (K)	i am not sure what you mean by " phpmyadmin <path>
Trans-Dec (K)	i am not sure what phpmyadmin is , but it 's probably in <path>
Trans-MLM (K)	i am not sure what you mean by " phpmyadmin - > <path> . <url>
Trans-AR (K)	i do not know what phpmyadmin is , but it 's in the repos
PF-free (K)	it ' s not in this server . i am trying to get phpmyadmin working
FG-free (K)	phpmyadmin is not in the same machine , it ' s just a server .
PF&FG-free (K)	i am not sure if it 's on the same machine , you can use phpmyadmin

Table 15: Responses generated by baselines and our methods trained on the **Ubuntu** dataset(million-scale and 100K).