DLGNet: A Transformer-based Model for Dialogue Response Generation

Oluwatobi Olabiyi
Capital One Conversation Research
Vienna VA
oluwatobi.olabiyi@capitalone.com

Erik T. Mueller
Capital One Conversation Research
Vienna VA
erik.mueller@capitalone.com

Abstract

Neural dialogue models, despite their successes, still suffer from lack of relevance, diversity, and in many cases coherence in their generated responses. On the other hand, transformer-based models such as GPT-2 have demonstrated an excellent ability to capture long-range structures in language modeling tasks. In this paper, we present DLGNet, a transformer-based model for dialogue modeling. We specifically examine the use of DLGNet for multi-turn dialogue response generation. In our experiments, we evaluate DLGNet on the open-domain Movie Triples dataset and the closed-domain Ubuntu Dialogue dataset. DLGNet models, although trained with only the maximum likelihood objective, achieve significant improvements over state-of-the-art multi-turn dialogue models. They also produce best performance to date on the two datasets based on several metrics, including BLEU, ROUGE, and distinct n-gram. Our analysis shows that the performance improvement is mostly due to the combination of (1) the long-range transformer architecture with (2) the injection of random informative paddings. Other contributing factors include the joint modeling of dialogue context and response, and the 100% tokenization coverage from the byte pair encoding (BPE).

1 Introduction

Recent successes of pretrained transformer-based language models, such as BERT (Devlin et al., 2019), GPT(-2) (Radford and Salimans, 2018; Radford et al., 2019), Transformer-XL (Dai et al., 2019), XLNet (Yang et al., 2019), and ERNIE(2.0) (Sun et al., 2019a,b), have led to state-of-the-art performance on many natural language understanding (NLU) tasks including sentence classification, named entity recognition, sentence similarity, and question answering. The exceptional performance of transformer-based language models is due to their ability to capture long-term temporal dependencies in the input sequence. This attribute should be very beneficial to dialogue modeling, especially in multi-turn scenarios. Most of the existing neural dialogue response generation models are based on recurrent neural networks (Sutskever et al., 2014; Vinyals and Le, 2015; Li et al., 2016a; Serban et al., 2016; Xing et al., 2017; Serban et al., 2017b,a; Li et al., 2016b; Zhang et al., 2018a; Olabiyi et al., 2018, 2019a).

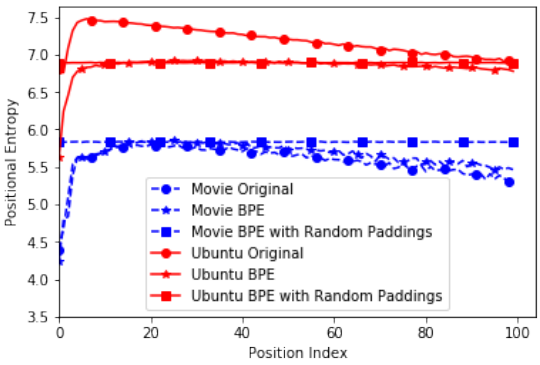

Figure 1: **Positional Entropy for Movie and Ubuntu datasets -** Applying a greedy training objective to the original and BPE datasets can achieve low overall entropy just by overfitting to low entropy regions, resulting in short and generic responses. Injecting random paddings into the data does not suffer from this problem and can be used to train transformer architectures due to their lack of recurrent propagations.

These models have yielded promising results by generating mostly coherent responses given the dialogue context. However, most of them, including the state-of-the-art models trained with naturalistic dialogue data, still perform well below the human level. Generated responses tend to be either generic, out-of-context, or disproportionately short. Previous work points to some causes of these limitations:

Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, pages 54–62

i) Training data: The presence of high frequency generic utterances (utterance-level semantic redundancy), such as "I don't know", "I'm not sure", and high frequency generic n-gram tokens (word-level syntactic redundancy), such as "I", "I am", leading to the concave positional entropy profile of dialogue datasets, see Fig. 1), which makes learning difficult, resulting in short and generic responses. *ii) Short-range Model Architecture:* Short-range model architectures that capture limited temporal dependencies. *iii) Out-of-vocabulary Problem:* Less frequent (usually more informative) words mapped to the out-of-vocabulary token <UNK>, leading to generation of a large number of <UNK> tokens. *iv) Exposure Bias:* The discrepancy in model behavior between training and inference, which limits the informativeness of the responses *iv) Training Objective:* The limitations of the maximum likelihood training objective.

In this paper, we propose DLGNet, a transformer-based model for multi-turn dialogue modeling that addresses some of the highlighted problems above. The use of a transformer architecture allows DLGNet to capture long-term temporal dependencies in the dialogue data better than the existing RNN-based architectures (Vaswani et al., 2017). However, applying a vanilla Seq2Seq transformer (Vaswani et al., 2017) and its multi-turn variants, such as ReCoSa (Zhang et al., 2019), for dialogue modeling does not work well because of the semantic redundancy in dialogue data. To overcome this, DLGNet models the joint distribution of the context and response instead of the conditional distribution of the response given the context, usually employed in Seq2Seq frameworks (Vinyals and Le, 2015; Serban et al., 2016; Olabiyi et al., 2018; Vaswani et al., 2017). DLGNet also addresses the syntactic redundancy in dialogue data by appending random paddings before and after the input data. This helps to break down the learning barrier from the concave entropy profile of human conversation data, as shown in Fig. 1. The flattening of the entropy profile also provides regularization during training, and reduces even the extent of the exposure bias problem. Finally, to avoid the out-of-vocabulary problem, DLGNet uses byte pair encoding (BPE) similar to GPT-2 (Radford et al., 2019) to provide 100% coverage for any Unicode input and output texts. Given all these proposed changes, we train DLGNet models using only the maximum likelihood objective. DLGNet models, despite being trained with only the maximum likelihood objective, demonstrate state-of-the-art performance on the Movie and Ubuntu datasets, as measured in terms of BLEU, ROUGE, and distinct n-gram scores.

Table 1: Example of generated responses from DLGNet models

Item	Utterance
Movie	
Context_0	says he wanted food . <person> . he wanted the gold .
Context_1	how ' s he going to want the gold ? he couldn ' t even know we had it .
Groundtruth	he . he could have been watching us .
DLGNet-117M [1]	he knew we did .
DLGNet-345M	<person> ' s just say he doesn ' t have enough to buy a trailer and seeds . i ' ll write him up as a decorated veteran of the united states army .
DLGNet-117M_Random [2]	<person> , you know the break-in , i might not be able to afford it .
DLGNet-345M_Random	are you going to preach now ?
Context_0	what are you thinking ?
Context_1	that i want you to take me home .
Groundtruth	now ? we just –
DLGNet-117M	now ? we ' re going now ? what about tomorrow ?
DLGNet-345M	now ? we ' re on the run !
DLGNet-117M_Random	now ?
DLGNet-345M_Random	no . keep me company . <person> offers its rewards . <person> appetit .
Ubuntu	
Context_0	The netboot one is suppose to download packages from the net.
Context_1	like the ones to be installed? or the installed to be run?
Groundtruth	Installed. The netbook also features the non-graphical installer.
DLGNet-117M	the installed to be run.
DLGNet-345M	the ones to be installed.
DLGNet-117M_Random	I think the netboot one is the one that is installed to the net.
DLGNet-345M_Random	the ones to be installed to.
Context_0	DJones: update manager won't detect 12.04.1 as a new version if you are already running 12.04, because 12.04.1 = 12.04 + lots of package updates
Groundtruth	DJones I am running 10.04.
DLGNet-117M	I'm not sure what you mean by "update manager won't detect"
DLGNet-345M	I see. I guess I'll just do a fresh install.
DLGNet-117M_Random	I'm running 12.04.1, and I'm not running 12.04.1.
DLGNet-345M_Random	I'm not sure what you mean by "update manager"

2 Task Description

Consider a dialogue sample consisting of a sequence of N utterances, $\mathbf{x} = \left(x_1, x_2, \cdots, x_N\right)$, where each utterance $x_i = \left(x_i^1, x_i^2, \cdots, x_i^{M_i}\right)$ contains a variable-length sequence of M_i word tokens such that ${x_i}^j \in V$ for vocabulary V. At any time step i, the dialogue history is given by $\mathbf{x}_i = \left(x_1, x_2, \cdots, x_i\right)$. The dialogue response generation task can be defined as follows: Given a dialogue history $\mathbf{x}_i$, generate a response $y_i = \left(y_i^1, y_i^2, \cdots, y_i^{T_i}\right)$, where T_i is the number of generated tokens such that the distribution of the generated response $P(y_i)$ is indistinguishable from that of the ground truth $P(x_{i+1})$. The distribution of the model output sequence can be factored by the

[1] Model with pretraining

[2] Model with random initialization (without pretraining)

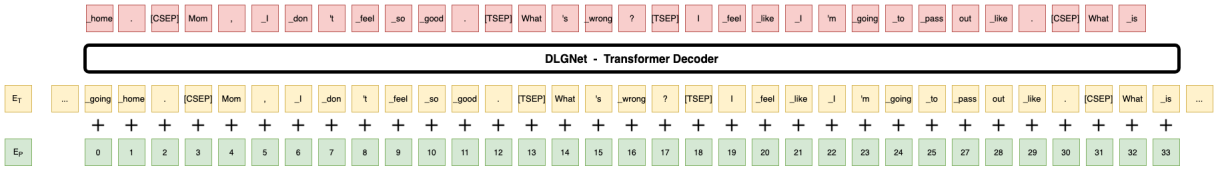

Figure 2: **An example of DLGNet input and output** consisting of a 3-turn conversation sample separated by [TSEP] tokens, combined with random informative paddings, before and after. Paddings and conversations are separated by [CSEP] tokens.

product rule:

$$P(y_i|\mathbf{x}_i) = \prod_{j=2}^{T_i} P\big(y_i^j|y_i^{1:j-1}, \mathbf{x}_i\big) \tag{1}$$

where $y_i^{1:j-1} = (y_i^1, \cdots, y_i^{j-1})$.

The MLE objective based on the conditional distribution of (1) can be expressed as

$$L_{\text{Cond}} = -\log P_\theta(y_i|\mathbf{x}_i) \tag{2}$$

where θ are the model parameters.

This formulation, known as Seq2Seq, originated from machine translation (Sutskever et al., 2014) and assumes that the context-response pair in the training examples are fairly unique. Seq2Seq is the basis of most of the previous work on dialogue modeling. The framework, however, does not account for the semantic and syntactic redundancy in human conversations as pointed out by Li et al. (2016a).

3 DLGNet Model Description

In order to address the semantic redundancy, we propose to jointly model both the context and the response as an alternative to the mutual information objective (Li et al., 2016a; Zhang et al., 2018b). The resulting distribution and the objective function can then be respectively expressed as

$$P(y_i, \mathbf{x}_i) = P(y_i|\mathbf{x}_i)P(\mathbf{x}_i) \tag{3}$$

$$L_{\text{Joint}} = -\log P_\theta(y_i|\mathbf{x}_i) - \log P_\theta(\mathbf{x}_i) \tag{4}$$

While (3) addresses the semantic redundancy, it does not address the syntactic redundancy coming from the concave positional entropy profile of dialogue data. To circumvent this, we append random informative paddings (sampled from the dataset) before ($\mathbf{x}_i^b$) and after ($\mathbf{x}_i^a$), the dialogue example of interest, leading to

$$P(\mathbf{x}_i^a, y_i, \mathbf{x}_i, \mathbf{x}_i^b) = P(\mathbf{x}_i^a)P(y_i|\mathbf{x}_i)P(\mathbf{x}_i)P(\mathbf{x}_i^b) \tag{5}$$

and

$$\begin{aligned} L_{\text{DLGNet}} = &-\log P_\theta(\mathbf{x}_i^a) - \log P_\theta(y_i|\mathbf{x}_i) \\ &- \log P_\theta(\mathbf{x}_i) - \log P_\theta(\mathbf{x}_i^b) \end{aligned} \tag{6}$$

since $\mathbf{x}_i^b$ and $\mathbf{x}_i^a$ are independent of $(y_i, \mathbf{x}_i)$. As we see from the resulting entropy profile in Fig. 1, appending random paddings circumvents the adverse effect of syntactic redundancy in dialogue data on model training. The conditional distribution $P(y_i|\mathbf{x}_i)$ in (1) is then just an inference on the joint distribution of (5).

DLGNet adopts GPT-2's autoregressive transformer architecture (Radford et al., 2019) using only the decoder part of the original transformer architecture (Vaswani et al., 2017) since there is no need for a separate encoder network (see Fig. 2). Autoregressive transformer models use multiple layers of masked multi-head self-attention to map a sequence of input tokens to a sequence of output tokens (i.e., the input sequence token shifted one position to the right). During inference, at each step, the model is autoregressive, consuming the previously generated token as additional input when generating the next. There are some basic conceptual differences between autoregressive architectures based on transformers and those based on recurrent neural networks (RNNs). For instance, while the output of an RNN layer depends on only the immediate previous output, a transformer layer output consists of attention over all previous outputs. Due to this lack of ordering in transformer architectures, the position representation is usually passed along with the input tokens into the model (Vaswani et al., 2017).

In order to take advantage and evaluate the impact of pretrained parameters, we use two model configurations i.e., (i) **DLGNet-117M** - with 117M parameters, 12 attention layers, and a hidden state size of 767, and (ii) **DLGNet-345M** - with 345M parameters, 24 attention layers, and a hidden state size of 1024; similar to the publicly available GPT-2 models (Radford et al., 2019).

Table 2: Automatic Evaluation of Model Performance

Model	Movie				Ubuntu			
	Relevance		Diversity		Relevance		Diversity	
	BLEU	ROUGE	DIST-1/2	NASL	BLEU	ROUGE	DIST-1/2	NASL
HRED	0.0474	0.0384	0.0026/0.0056	0.535	0.0177	0.0483	0.0203/0.0466	0.892
VHRED	0.0606	0.1181	0.0048/0.0163	0.831	0.0171	0.0855	0.0297/0.0890	0.873
hredGAN_u	0.0493	0.2416	0.0167/0.1306	0.884	0.0137	0.0716	0.0260/0.0847	1.379
hredGAN_w	0.0613	0.3244	0.0179/0.1720	**1.540**	0.0216	0.1168	0.0516/0.1821	1.098
DAIM	0.0155	0.0077	0.0005/0.0006	0.721	0.0015	0.0131	0.0013/0.0048	**1.626**
aBoots_u_cat	0.0880	0.4063	0.0624/0.3417	0.918	0.0210	0.1491	0.0523/0.1795	1.040
aBoots_w_cat	0.0940	0.3973	0.0613/0.3476	1.016	0.0233	0.2292	0.1288/0.5190	1.208
DLGNet-117M_Random	0.1796	0.4338	0.1198/0.4578	1.011	0.0215	0.1978	0.1827/0.4074	0.829
DLGNet-345M_Random	0.2682	0.4881	0.1286/0.4612	0.907	**0.0315**	0.2041	0.1927/0.4468	0.794
DLGNet-117M	0.1872	0.4346	0.1232/0.4506	0.982	0.0279	0.2191	0.2228/0.4953	0.746
DLGNet-345M	**0.2742**	**0.4945**	**0.1282/0.4736**	0.895	**0.0309**	**0.2409**	**0.2436/0.5632**	0.759

4 Model Training

We trained the small DLGNet-117M and the medium DLGNet-345M models on multi-turn dialogue datasets initialized with either random noise or pretrained language model parameters. The models are trained end-to-end using the Adaptive Moment Estimation (Adam) stochastic gradient descent algorithm with a learning rate of 0.001. The maximum sequence length is 1024. Due to GPU memory limitations, we use a batch size of 2 and accumulate gradients over 5 iterations, making the effective batch size 10. Both models are trained until the training perplexity on the dialogue datasets reaches a steady state. Finally, the models are implemented, trained, and evaluated using Python and the TensorFlow deep learning framework.

5 Experiments

5.1 Setup

We evaluated DLGNet models on the Movie Triples and Ubuntu Dialogue corpora randomly split into training, validation, and test sets, using 90%, 5%, and 5% proportions. Since we use BPE with 100% tokenization coverage, we performed no preprocessing of the datasets whatsoever. For each training example, however, we randomly sample a target conversation and two independent padding chunks from the dataset to fill up the maximum input sequence length. We append the paddings to the target conversation, one before, and one after, separated by token [C_SEP]. The target conversation in each training example in turn consists of utterances that are separated by token [T_SEP] as shown in Fig. 2.

The Movie dataset (Serban et al., 2016) spans a wide range of topics with few spelling mistakes and contains about 240,000 dialogue triples, which makes it suitable for studying the relevance-diversity tradeoff in multi-turn conversations (Zhang et al., 2018b). The Ubuntu dialog dataset extracted from the Ubuntu Relay Chat Channel (Serban et al., 2017b) contains about 1.85 million conversations with an average of 5 utterances per conversation. This dataset is ideal for training dialogue models that can provide expert knowledge/recommendation in domain-specific conversations.

We compare DLGNet multi-turn dialogue performance with existing state-of-the-art dialogue models including (V)HRED[3] (Serban et al., 2016, 2017b), DAIM[4] (Zhang et al., 2018b), hredGAN (Olabiyi et al., 2018), and aBoots (Olabiyi et al., 2019b). Note that DAIM is single turn and does not use a multi-turn dialogue context, but we have included it here for completeness. We compare how the models perform based on informativeness (a combination of relevance and diversity metrics) of generated responses. For relevance, we adopted BLEU-2 (Papineni et al., 2002) and ROUGE-2 (Lin, 2014) scores. For diversity, we adopted distinct unigram (DIST-1) and bigram (DIST-2) (Li et al., 2016a) scores as well as normalized average sequence length (NASL), similar to Olabiyi et al. (2018).

All models are evaluated in autoregressive mode, i.e., we pass a multi-turn dialogue context to the model inputs and the models generate a sequence of response tokens using the context and all the previously generated tokens until the end-of-sequence

[3]implementation obtained from https://github.com/julianser/hed-dlg-truncated

[4]implementation obtained from https://github.com/dreasysnail/converse_GAN

token is reached. All models are greedily sampled to generate the model outputs. It is worth noting that, for DLGNet models, we search for the optimum top_k between 0 and 20 inclusive that maximizes the overall BLEU-2 (relevance) score of the validation set using the top_k sampling strategy (Radford et al., 2019). It turns out that for all DLGNet models, the optimum top_k is 1 across datasets, which is equivalent to greedy sampling.

6 Results and Discussion

6.1 Quantitative Evaluation

We report the quantitative measures in Table 2. The transformer-based DLGNet provides a significant improvement in response generation performance over existing methods such as (V)HRED, hredGAN, DAIM, and adversarial bootstrapping (aBoots), all of which are based on recurrent neural networks. In fact, DLGNet achieves the best performance to date on the Movie triples and Ubuntu dialogue datasets in terms of BLEU, ROUGE, and distinct n-gram scores. This indicates that, despite being trained only with the maximum likelihood objective, the autoregressive transformer architecture in conjunction with the random padding injection, is able to overcome some of the problems that have plagued existing dialogue models such as semantic and syntactic redundancy, and exposure bias. Also contributing to the models' performance improvement is the 100% input coverage from the BPE encoding, which eliminates the generation of <UNK> tokens (this is especially helpful for the Ubuntu dataset with a large number of out-of-vocabulary tokens) as well as the joint modeling of the context and response. Also, in contrast to existing work reporting a trade-off between relevance and diversity (Zhang et al., 2018b; Li et al., 2016a,b), we observe that relevance performance improves with diversity performance in DLGNet models. It is worth pointing out, however, that DLGNet models tend to generate shorter responses than adversarially trained models (hredGAN and aBoots). This indicates that the models still suffer from the impact of using only the maximum likelihood training objective. Alleviating this problem with an adversarial training objective similar to aBoots and or hredGAN should further improve performance and will be considered in our future work.

6.2 Qualitative Evaluation

Random samples of the model outputs are shown in Tables 1 and 4. One striking observation is the high level of coherence in the generated responses from DLGNet models. The models are able to capture both short- and long-term temporal dependencies in their responses. The models give responses that are relevant to the topic of the discussion, and are able to answer posed questions with answer choices. Also, they don't simply generate the all-too-common phrase "I'm not sure" like existing models; they are able to point to areas of the context they are uncertain about (see the Ubuntu section of Table 1).

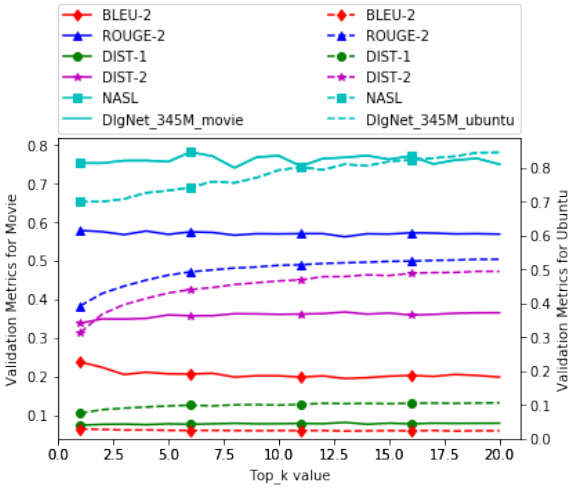

Figure 3: Relevance vs. diversity tradeoff with top_k sampling for DLGNet-345M models.

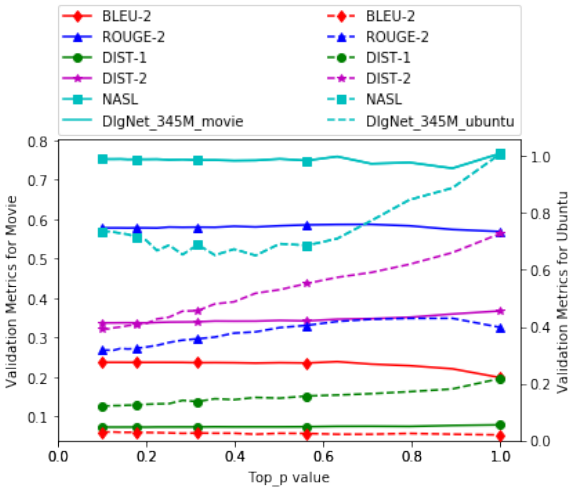

Figure 4: Relevance vs. diversity tradeoff with top_p sampling for DLGNet-345M models.

7 Ablation Studies on DLGNet Models with Random Informative Padding

In this section, we carry out a more detailed analysis and discussion of different configurations of DLGNet models as well as their performance across datasets, using the evaluation results in Table 2.

7.1 Open vs. Closed Domain Dataset

From Table 2, we observe that the performance improvement achieved by DLGNet models over existing models is higher for the open-domain Movie Triples dataset than for the closed-domain Ubuntu Dialogue dataset with or without pretraining. While the performance difference could be due to the size of the dataset, it could also indicate that closed-domain dialogue responses are inherently more difficult to learn, even for large and expressive models such as the DLGNet transformer.

7.2 Effect of Model Pretraining

Although models with pretraining generally perform better than ones trained with random initialization, we observe that the performance difference is not significant. This shows that the performance of the DLGNet is mostly due to the multi-layer self attention model architecture rather than the scaffolding achieved from language model pretraining. We observe similar behavior across datasets. However, pretraining seems to be consistently more helpful for open-domain datasets versus closed-domain datasets. This might be because the distribution of the language data used for pretraining is similar to the open-domain dataset but different from the closed-domain dataset. Also, models without pretraining tend to generate longer responses on average compare to those with pretraining. This indicates that model pretraining also plays a role in the relevance-diversity tradeoff.

7.3 Effect of Model Size

We also compare the small (DLGNet-117M) and large (DLGNet-345M) models. We observe that there is a significant performance improvement of the larger over the smaller model on the Movie dataset (about 50%), but a smaller performance improvement on the Ubuntu dataset. It's also surprising that the larger model doesn't overfit to the Movie dataset. Overfitting might have been prevented by the injection of random padding into the input data, which regularizes the model training by artificially inducing high entropy into the data.

7.4 Relevance vs. Diversity Tradeoff

The results in Table 2 show state-of-the-art relevance performance with some compromise on the response length. Here, we explore the possibility of generating longer and more diverse responses with the trained models and estimate the effect on the relevance scores. For this experiment, we chose the larger DLGNet-345M models of both datasets and tried two sampling techniques, i.e., top_k (Radford et al., 2019) and top_p nucleus (Holtzman et al., 2019; Zellers et al., 2019) sampling strategies on the validation sets. The trajectory of the evaluation metrics with increasing top_k and top_p values are shown Figs. 3 and 4 respectively. With top_k sampling, increasing the top_k value increases the response length at the expense of relevance metrics like BLEU for both datasets, as expected. However, the response length increase is more significant on the Ubuntu dataset than the Movie dataset. It is also surprising that the ROGUE-2 score for Ubuntu increases with increasing top_k value, which is the reverse of the case for the Movie dataset. Also, Fig. 3 shows that it is more advantageous to trade off relevance for diversity on the Ubuntu dataset compare to the Movie dataset. This is probably due to the size and closed-domain nature of the Ubuntu dataset, which makes it more difficult to learn with the maximum likelihood estimation only.

We observe a similar pattern with the top_p nucleus sampling in Fig. 4. This reinforces the fact that greedy sampling may be sufficient for open-domain datasets such as Movie.

8 Further Ablation Studies on DLGNet Models

We also set out to analyze the features of DLGNet that make it suitable for multi-turn dialogue modeling. We train both DLGNet-117M and DLGNet-345M models on both datasets, but replace the random informative paddings with static paddings using a pad token. Below are the definitions of the model configuration factors considered:

1.) Multi-turn Data (M): Training data is variable-length multi-turn data padded to a fixed length. This helps to evaluate the effect of using random informative padding.

2.) Single-turn Data (S): Training data is variable-length single-turn data padded to a fixed length. This helps to evaluate the effect of number of turns.

3.) Joint model (Joint): DLGNet models are

Table 3: Ablation Performance of DLGNet Models with Static Padding

Model	Movie				Ubuntu			
	Relevance		Diversity		Relevance		Diversity	
	BLEU	ROUGE	DIST-1/2	NASL	BLEU	ROUGE	DIST-1/2	NASL
DLGNet-117M								
S-Joint with BPE	∼0.0	∼0.0	0.0400/0.1502	0.072	∼0.0	0.0004	0.1946/0.4636	0.064
S-Cond with BPE	0.0013	0.0296	0.0134/0.0482	3.582	∼0.0	0.0083	0.0723/0.1470	0.890
M-Joint with BPE	0.1825	0.1321	0.0346/0.0838	0.610	0.0012	0.1172	0.1719/0.3482	0.2937
M-Cond with BPE	0.0096	0.0628	0.0088/0.0394	3.425	0.0048	0.0766	0.0500/0.1454	2.372
M-Joint with Basic Tokenizer	0.0518	0.0630	0.0176/0.0540	1.101	0.0030	0.0384	0.0465/0.0949	0.566
M-Cond with Basic Tokenizer	0.0149	0.1628	0.0394/0.1770	1.472	∼0.0	0.0136	0.2211/0.4192	0.281
DLGNet-345M								
S-Joint with BPE	∼0.0	∼0.0	∼0.0/∼0.0	0.072	∼0.0	0.0006	0.4741/0.9760	0.061
S-Cond with BPE	0.0006	0.0212	0.0010/0.0419	3.582	0.0004	0.0158	0.0721/0.1671	3.437
M-Joint with BPE	0.0449	0.1931	0.0460/0.1273	0.531	∼0.0	0.0121	0.3323/0.4406	0.227
M-Cond with BPE	0.0010	0.0125	0.0091/0.0422	3.918	0.0004	0.0158	0.0721/0.1671	4.108
M-Joint with Basic Tokenizer	0.0376	0.1389	0.0232/0.0654	0.543	0.0042	0.0341	0.0568/0.1299	0.552
M-Cond with Basic Tokenizer	0.0057	0.0970	0.1568/0.3785	0.331	0.0015	0.0345	0.1555/0.3990	0.470

trained by jointly modeling the dialogue context and response.

4.) Conditional model (Cond): DLGNet models are trained in the traditional sequence-to-sequence mode with a bidirectional encoder and an autoregressive decoder for a conditional modeling of the dialogue response given the context (Vaswani et al., 2017; Zhang et al., 2019).

5.) Basic Tokenizer: We use a basic tokenization traditionally used in dialogue modeling instead of BPE tokenization to evaluate the effect of tokenization coverage. It also provides an apples-to-apples comparison between the transformer-based and RNN-based architectures.

8.1 Effect of Random Padding Injection

The results in Table 3 are from models trained with static paddings. The models perform significantly worse than those of Table 2. Without random padding injection, the models quickly overfit to the low entropy regions of the training data, which leads generic and/or short responses.

8.2 Single Turn vs. Multi-turn

We also observe that the multi-turn models perform better than single-turn models on BPE tokenized data. This is expected because the multi-turn models capture longer temporal dependencies in the input data. It is also worth mentioning that the single-turn performance is further hurt by BPE tokenization since it tends to work better with long input sequences.

8.3 Joint vs. Conditional Models

For multi-turn models, the joint modeling architecture yields better performance than the conditional Seq2Seq architecture. This trend is however reversed for single-turn models. This is because a model that focuses on jointly modeling both the context and the response performs better with longer contextual information compared to a model that focuses on modeling only the conditional distribution of the response given the context. Therefore, multi-turn dialogue model should rather employ the joint structure instead of the conditional Seq2Seq structure.

8.4 Effect of Tokenization Coverage

For a more fair comparison with previous work on multi-turn dialogue not using random padding injection and 100% BPE tokenization, we trained the DLGNet models on multi-turn data with basic tokenization. The tokenization coverages of the basic tokenizer used are 83.9% and 4.19% for Movie and Ubuntu datasets respectively. Basically, most of the Ubuntu tokens are mapped to the <UNK> token. In comparison with previous work on HRED, the results in Table 3 show that the transformer-based DLGNet models under the same conditions perform better than the basic HRED model but worse than the improved HRED models (such as VHRED, hredGAN, and aBoots). In comparison with other transformer-based configurations, the smaller size multi-turn models perform better than their BPE counterparts but the larger size models perform worse. This is probably due to the overfitting of the larger models.

9 Conclusion

In this paper, we have proposed DLGNet, an extension of autoregressive transformer models such as GPT-2 for multi-turn dialogue modeling. Our experiments show that DLGNet models perform better than existing state-of-the-art multi-turn dialogue models. They also achieve the best performance to date on open-domain Movie and closed-domain Ubuntu datasets based on BLEU, ROUGE and distinct n-gram scores. Our experiments reveal that the combination of (i) the transformer architecture with (ii) the injection of random paddings exploiting the large maximum input sequence is responsible for the performance improvement over existing methods. Other contributing factors include joint modeling of dialogue context and response, and the 100% tokenization coverage from the byte pair encoding (BPE). Our analysis also reveals some tradeoffs between response relevance and response length, and we showed how different sampling strategies can be used to make an informed decision about such response relevance-diversity compromises. In our future work, we plan to investigate how to improve on the length of the generated responses without necessarily sacrificing their coherence and their relevance to the dialogue context.

Table 4: Sample responses of DAIM, (V)HRED, hredGAN, aBoots and DLGNet models.

Model **Movie**	**Response**
Context_0	we ' re grateful , <person> . it ' s not that . those craftsman tools . they ' re a dream come true .
Context_1	your dad and i were raised with a certain work ethic , that ' s all .
Groundtruth	we can ' t start living like we ' re the rockerfellers or something .
HRED	<person> .
VHRED	<person> , mr . <person> . <person> , i ' m not going to tell you , <person> . i ' m sorry . i ' m sorry , <person> . i ' m sorry
hredGAN_w	<person> , i know you ' re see . < continued_utterance > <person> , please .
aBoots_w_cat	you ' re a liar !
DLGNet-117M	<person> , come on , hang on . how do you like that ? i can barely believe you ' re doing this .
DLGNet-345M	i have some very bad dreams .
DLGNet-117M_Random	i ' m sorry about all this . <person> , this is all their house .
DLGNet-345M_Random	remember that shoe you lost .
Context_0	<person> i ' ve raised my rate . i might even cut my sessions to forty minutes .
Context_1	<person> , i know you don ' t like flattery but if anybody i know is going to win a nobel prize , it ' s you . you got ta be thinking about your next book so i know you ' ll find this case particularly interesting .
Groundtruth	what sort of case is it , <person> ?
HRED	i ' m sorry , < person> .
VHRED:	it ' s not a lot of <unk> . i ' m not going to tell you , <person> . i ' m not going to help you .
hredGAN_w	i ' m a little nervous , but i don ' t care if there ' s something more important than <person> .
aBoots_w_cat	if i ' m sure they ' ll call it ' s place <person> .
DLGNet-117M	what sort of case is it , <person> ?
DLGNet-345M	what sort of case is it , <person> ?
DLGNet-117M_Random	what sort of case is your next ?
DLGNet-345M_Random	like what ?
Context_0	what was that for ?
Context_1	i ' d forgotten how exciting it is working together .
Groundtruth	you ' ve been so quiet lately , <person> , i thought you no longer cared .
HRED	i ' m sorry , <person> .
VHRED:	<person> , <person> . i don ' t know .
hredGAN_w	i ' m sorry , <person> . so many people are so interested , but you ' re a very nice guy .
aBoots_w_cat	we ' re going to take a look ?
DLGNet-117M	you look like working with <person> .
DLGNet-345M	you ' re a helluva translator .
DLGNet-117M_Random	you ' ve been so nice to <person> ach .
DLGNet-345M_Random	oh .
Ubuntu	
Context_0	DJones: update manager won't detect 12.04.1 as a new version if you are already running 12.04, because 12.04.1 = 12.04 + lots of package updates
Context_1	DJones I am running 10.04.
Context_2	what does 'gksu update-manager c' report?
Groundtruth	does not report that there is a new version out. So I am just holding. No rush.
HRED	I don't know what the problem is
VHRED	you mean ?
hredGAN_w	yes , I get the new one , I just want to get the new one
aBoots_w_cat	it says : no such file or directory , that is what it says
DLGNet-117M	gksu: command not found
DLGNet-345M	it says No new release found
DLGNet-117M_Random	http://pastebin.com/m7e8e0e7c
DLGNet-345M_Random	it says "update-manager: error: no new version is available"
Context_0	libgnome2-dev is the package that has that pkgconfig file
Context_1	After I installed libgnome2-dev I got a different output. http://pastebin.com/zpDeGs81 sorry it is so long.
Context_2	I thought that was what was going on thanks. :) I ran the command ./configure && sudo make && sudo make install would that have worked.
Groundtruth	you shouldn't run: sudo make; but yes
HRED	I don't know what the problem is
VHRED	what do you mean ? __eou__ what are you trying ?
hredGAN_w	I have no idea , I can just copy the files and copy files to the other one
aBoots_w_cat	yeah , if you have a few times to put things in the ' make ' command it is not in a file you can install it by hand .
DLGNet-117M	no, that's not what you want to do
DLGNet-345M	no, you need to run it as root
DLGNet-117M_Random	no, it won't.
DLGNet-345M_Random	yes, that's what I did

References

Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

A. Holtzman, J. Buys, M. Forbes, and Y. Choi. 2019. The curious case of neural text degeneration. In *arXiv preprint arXiv:1904.09751*.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*.

J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. 2016b. Deep reinforcement learning for dialogue generation. In *arXiv preprint arXiv:1606.01541v4*.

C. Y. Lin. 2014. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*.

O. Olabiyi, A. Khazan, A. Salimov, and E.T. Mueller. 2019a. An adversarial learning framework for a persona-based multi-turn dialogue model. In *NAACL NeuralGen Workshop*.

O. Olabiyi, E.T. Mueller, C. Larson, and T. Lahlou. 2019b. Adversarial bootstrapping for dialogue model training. In *arXiv preprint arXiv:1909.00925*.

O. Olabiyi, A. Salimov, A. Khazane, and E. Mueller. 2018. Multi-turn dialogue response generation in an adversarial learning framework. In *arXiv preprint arXiv:1805.11752*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: A method for automatic evalution of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

A. Radford and T. Salimans. 2018. Improving language understanding by generative pre-training. In *https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf*.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. In *https://d4mucfpksywv.cloudfront.net/better-language-models*.

I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of The Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*, pages 3776–3784.

I. V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. Courville. 2017a. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of The Thirty-first AAAI Conference on Artificial Intelligence (AAAI 2017)*.

I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogue. In *Proceedings of The Thirty-first AAAI Conference on Artificial Intelligence (AAAI 2017)*.

Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu. 2019a. Ernie: Enhanced representation through knowledge integration. In *arXiv preprint arXiv:1904.09223*.

Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang. 2019b. Ernie 2.0: A continual pre-training framework for language understanding. In *arXiv preprint arXiv:1907.12412*.

I. Sutskever, O. Vinyals, and Q. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *NIPS*.

O. Vinyals and Q. Le. 2015. A neural conversational model. In *Proceedings of ICML Deep Learning Workshop*.

C. Xing, W. Wu, Y. Wu, M. Zhou, Y. Huang, and W. Ma. 2017. Hierarchical recurrent attention network for response generation. In *arXiv preprint arXiv:1701.07149*.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *arXiv preprint arXiv:1906.08237*.

R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2019. Defending against neural fake news. In *arXiv preprint arXiv:1905.12616*.

H. Zhang, Y. Lan, L. Pang, J. Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *arXiv preprint arXiv:1907.05339*.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *arXiv preprint arXiv:1801.07243v3*.

Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*.