Adaptive Bridge between Training and Inference for Dialogue Generation

Haoran Xu[1,2][*]**, Hainan Zhang**[3][†]**, Yanyan Zou**[3]**, Hongshen Chen**[3]**,**
Zhuoye Ding[3]**, Yanyan Lan**[4]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3]Data Science Lab, JD.com, Beijing, China
[4]Institute of AI Industry Research, Tsinghua University, Beijing, China
`xuhaoran18s@ict.ac.cn,zhanghainan1990@163.com,zouyanyan6@jd.com`
`ac@chenhongshen.com,dingzhuoye@jd.com,lanyanyan@tsinghua.edu.cn`

Abstract

Although exposure bias has been widely studied in some NLP tasks, it faces its unique challenges in dialogue response generation, the representative one-to-various generation scenario. In real human dialogue, there are many appropriate responses for the same context, not only with different expressions, but also with different topics. Therefore, due to the much bigger gap between various ground-truth responses and the generated synthetic response, exposure bias is more challenging in dialogue generation task. What's more, as MLE encourages the model to only learn the common words among different ground-truth responses, but ignores the interesting and specific parts, exposure bias may further lead to the common response generation problem, such as "I don't know" and "HaHa?" In this paper, we propose a novel adaptive switching mechanism, which learns to automatically transit between ground-truth learning and generated learning regarding the word-level matching score, such as the cosine similarity. Experimental results on both Chinese STC dataset and English Reddit dataset, show that our adaptive method achieves a significant improvement in terms of metric-based evaluation and human evaluation, as compared with the state-of-the-art exposure bias approaches. Further analysis on NMT task also shows that our model can achieve a significant improvement.

1 Introduction

Auto-regressive models(ARM) are widely used for natural language generation(NLG) tasks, such as machine translation (Sutskever et al., 2014; Wu et al., 2018), dialogue response generation (Li et al., 2017), image captioning (Lin et al., 2014; Vinyals et al., 2015) and video description (Donahue et al., 2015). They utilize the encoder-decoder framework to predict the next token conditioned

Dialogue 1	
context	听说广州已成避暑胜地
	(I heard that Guangzhou has become a summer resort)
response1	确实，这边很凉快。(Indeed, it's cool here.)
response2	晚上睡觉都没开风扇了。
	(There is no need to turn on the fan at night.)
Dialogue 2	
context	哈哈，看看可爱的小猫咪(Ha ha, look at this lovely kitten)
response1	这是什么品种的猫哇? 好可爱，我也想要
	(What kind of cat is this? So cute, I want it too.)
response2	好想要一只这样的猫，可以陪我儿子玩
	(I really want a cat like this to play with my son.)
response3	哇，好可爱哇，我屋也有一只这样的小猫
	(Wow, It's so cute, I have a kitten like this in my house.)

Table 1: The two Dialogues in STC dataset, and the red part of responses in Dialogue 2 are the common words.

on the previous tokens, and minimize the cross-entropy between the generation and ground-truths as their objective function. Specifically, at training time, the ground-truth is utilized as the previous tokens, which forces the model directly to learn the distribution of ground truths. But at inference, the previous tokens come from the ARM decoder itself, which is different from the input distribution at training time.

Although this discrepancy, named exposure bias, has been studied in some classic NLG tasks, such as neural machine translation(NMT) (Bengio et al., 2015; Venkatraman et al., 2015; Zhang et al., 2019), it faces its unique challenges in dialogue response generation, the representative one-to-various generation scenario. In human dialogue, given the context, people can reply many relevant and appropriate responses, not only with various expressions but also with different topics. Take the Dialogue 1 in Table 1 as an example, given the context "I heard that Guangzhou has become a summer resort", the response 1 and response 2 are in the same topic but with different tokens. In this various expression situation, like NMT task, data distribution and model distribution are easy to fit, relatively, even with exposure bias problem. However, in different topics situation, data distribution is often

---

different from the model, because it is too divergent and covers various word distribution of each topic. Through our data analysis, we find that in dialogue generation task, the various ground-truth responses and the generated sentences have a bigger gap than in NMT tasks. We calculate the overlap measures at word-level and semantic-level, i.e., BLEU and cosine similarity, between the generated sentence and the ground-truth sentences. The results show that on NMT WMT'14 dataset, the BLEU and similarity are 27.38 and 0.96, respectively, while on dialogue Reddit dataset, the BLEU and similarity are 2.17 and 0.81, respectively. We can see that the overlap measures of the dialogue generation task are significantly lower than that of the NMT task, which indicates the severity of the exposure bias problem in the dialogue generation.

What's more, as Maximum Likelihood Estimation(MLE) encourages the model to only learn the common words among different ground-truth responses, but ignores the interesting and specific parts, exposure bias may aggravate the common response problem of the generation model, due to the strict matching between the generated response and the ground-truth responses. Take the Dialogue 2 in Table 1 as an example, the response 1 is "What kind of cat is this? So cute, I want it too.", the response 2 is "I really want a cat like this to play with my son. " and the response 3 is "Wow, it's so cute, I have a kitten like this in my house.". If we train the model with word-level strict matching between the generated response and the ground-truth, it can only learn the common words, i.e., "So cute, I want it", but ignore the specific parts, i.e., "What kind of cat is this?". Therefore, it is beneficial to improve the strict matching mechanism for the dialogue generation task.

In this paper, we propose a novel Adaptive switch mechanism as a Bridge(AdapBridge), which introduces the generator distribution to the training phase and learns to automatically transit between the ground-truth learning and the generated learning, with respect to the word-level matching scores, such as the cosine similarity. Specifically, at each training step, we calculate the cosine similarity for each generated word with respect to all its ground-truths. If the matching score is bigger than the threshold, the generated word is fed to the decoder, while if lower, the ground-truth is fed for training. The threshold is increasing as the training epoch grows. With this adaptive sampling scheme,

the switch mechanism can consider the generation quality of every word, i.e., relevance between the generated word and the ground-truth, to decide whether utilizing the generated learning or not.

We evaluate the proposed models on two public datasets, i.e. the Chinese STC and the English Reddit dataset. Experimental results show that our models significantly outperform the state-of-the-art exposure bias models with respect to both metric-based evaluations and human judgments. Further analysis on NMT task also shows that our model can achieve a significant improvement.

The main contributions of this paper include:

- We study the exposure bias problem in dialogue generation task, one of the one-to-various generation scenarios. And find that the exposure bias may further lead to the common response generation problem.

- We propose the adaptive switch mechanism with word-level matching scores to determine the training input source, in order to resolve the common response problem.

- We evaluate AdapBridge on two public dialogue datasets and conduct rigorous experiments to demonstrate the effectiveness of our proposed models. Further analysis on NMT task also shows that our model can achieve a significant improvement.

2 Related Work

This section briefly introduces recent research progresses related to this work in literature.

To solve the exposure bias problem in auto-regressive or seq2seq models (Sutskever et al., 2014; Welleck et al., 2019; Holtzman et al., 2019), Venkatraman et al. (2015) tried to use data as Demonstrator(DAD) to augment the training set through the tokens predicted by the model, so as to make the training set to meet the test distribution. The method of Scheduled Sampling(SS) proposed by Bengio et al. (2015) attempted to randomly sample the previous generated words to replace the ground-truth words for the model input during training time. Zhang et al. (2019) made a further exploration of this method by sampling the previous words with decay not only from word-level oracle but also from the sentence-level oracle with a semantic metric. The main idea of this kind of method is to introduce the model's prediction information to its input at training time, and reduce
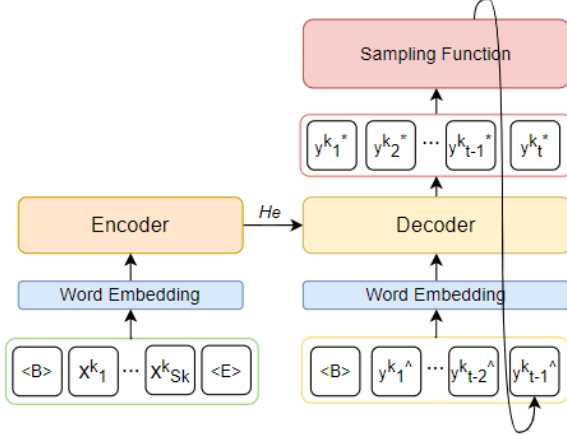
Figure 1: The illustration of our AdapBridge Model.

the discrepancy between training and inference to alleviate the exposure bias problem. In comparison to those methods and related ideas (Qi et al., 2020; Goodman et al., 2020), our proposed method adaptively determines whether the input words of model during training are ground truth or predicted by scoring each generated word.

Alternative based on Reinforcement Learning(RL) (Williams, 1992) methods have been explored for generation tasks, in particular for NMT. Mixed Incremental Cross-Entropy Reinforce (MIXER) (Ranzato et al., 2016) leverage hybrid loss function which combines both cross-entropy and reinforce to directly optimized the metrics used at test time, such as BLEU or ROUGE. There are many other similar works (Shen et al., 2016; Wu et al., 2016; Shao et al., 2018). More recently, text generation via Generative Adversarial Networks(GAN) (Goodfellow et al., 2014) called Text GANs has attracted of researchers (Nie et al., 2019; Zhou et al., 2019; Wu et al., 2021; Scialom et al., 2020). They framed the problem under the GAN paradigm, which uses the RL-Based (Williams, 1992) algorithms to get the gradient estimation, as the text generation is discrete. However, both RL and Text GANs cannot be avoided the high variance of gradient estimation caused by sparse rewards, which consequently makes the training process unstable and limits improvements.

Different from traditional methods, our proposed model can adaptively determine whether the current input word is from ground truth or from generation with the word-level matching scores.

3 Proposed Method

Given a context sentence $X^k = \{x_1^k, x_2^k, \cdots, x_{S_k}^k\}$, and a target response sentence $Y^k = \{y_1^k, y_2^k, \cdots, y_{T_k}^k\}$, where S_k and T_k are the word length of context and response, respectively. The dialogue generation model based on sequence-to-sequence (Seq2Seq) (Sutskever et al., 2014) framework, directly models the response probability:

$$P(Y^k|X^k, \theta) = \prod_{t=1}^{T_k} p(y_t^k|y_{<t}^k, X^k, \theta) \quad (1)$$

where θ are the parameters of model and $y_{<t}^k$ denote the previous ground-truth words. Given a set of training examples $D = \{X^k, Y^k\}_{k=1}^N$, the stander training objective is to minimize the negative log-likelihood of all the training data:

$$\theta = \underset{\theta}{argmin}\{L(\theta)\} \quad (2)$$

where

$$L(\theta) = \sum_{k=1}^{N} -logP(Y^k|X^k, \theta)$$
$$= \sum_{k=1}^{N} \sum_{t=1}^{T_k} -logp(y_t^k|y_{<t}^k, X^k, \theta) \quad (3)$$

where N is the number of training examples.

Different from training time, during inference, the probability of each target word $p(y_t^k|y_{<t}^k, X^k, \theta)$ in Equation 1 is conditioned on the previous generated words $y_{<t}^k{}^*$ rather than the ground-truth $y_{<t}^k$, as the ground truth words are not available in real inference time. This discrepancy is called exposure bias.

3.1 AdapBridge

The architecture of our model is illustrated in Figure 1. We first define a sampling function:

$$y_{<t}^k{}^{\wedge} = \mathbf{SamFun}(y_{<t}^k, y_{<t}^k{}^*) \quad (4)$$

where $y_{<t}^k$ and $y_{<t}^k{}^*$ are the inputs of sampling function, representing the ground truth words and generated words, respectively, and $y_{<t}^k{}^{\wedge} = \{y_1^k{}^{\wedge}, y_2^k{}^{\wedge} \cdots, y_{t-1}^k{}^{\wedge}\}$ denotes the inputs of decoder after sampled by the sampling function at $t-th$ time step, which may contain both ground truth and generated words.

In this framework, to predict $t-th$ target word $y_t^k{}^*$, we can follow those steps:
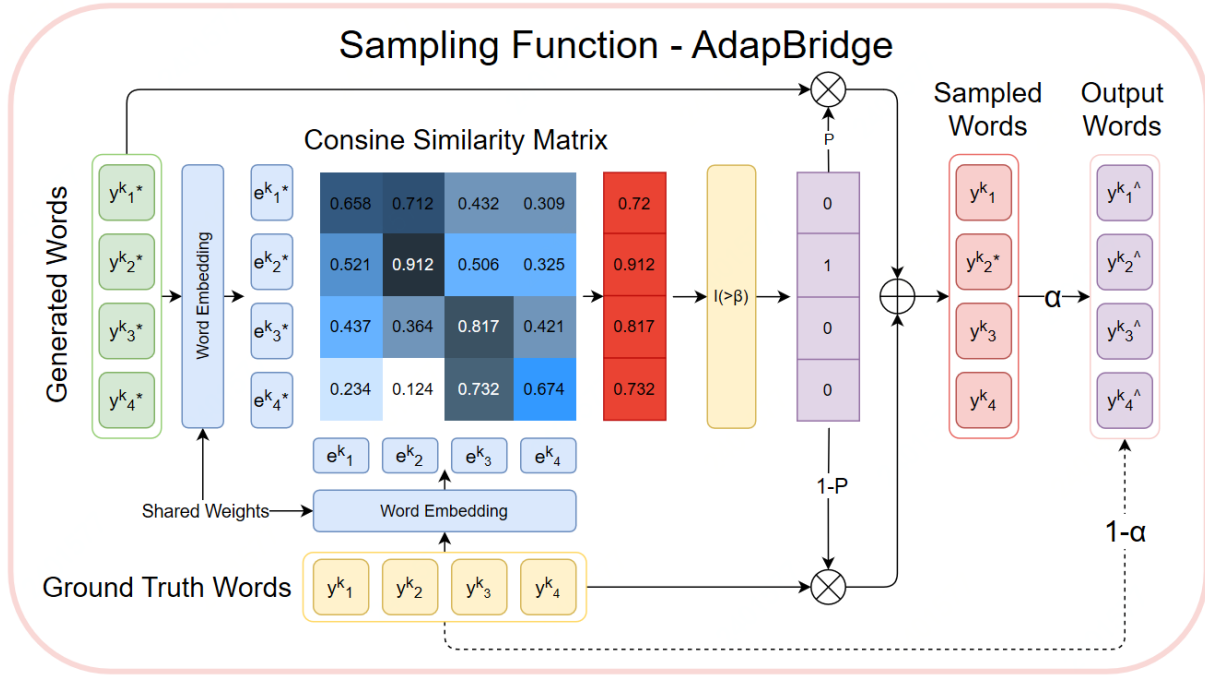
Figure 2: The Illustration of the sampling function Adapt-Bridge. The word embedding shares weights with encoder and decoder of model showed in Figure 1. $I(> \beta)$ is a Indicator function, if input is upper β, output will be 1, otherwise, output will be 0. α and β are both increasing as the training epochs grows.

- Decoders predict $t - 1$ words $y_{<t}^{k}{}^{*}$ as the previous generated words.

- Use the sequences $y_{<t}^{k}{}^{*}$ and $y_{<t}^{k}$(ground truth words) as inputs of **SamFun**(Equation 4), and get the outputs $y_{<t}^{k}{}^{\wedge}$ of this function.

- Replace the inputs $y_{<t}^{k}$ of model in Equation 1 with $y_{<t}^{k}{}^{\wedge}$, then predict the t-th word y_{t}^{k*}.

The **SamFun** can be any function, i.e. random sampling. The process is to replace the corresponding ground-truth words in $y_{<t}^{k}$ with the generated words $y_{<t}^{k}{}^{*}$. We propose a novel **SamFun** called AdapBridge, which can be seen in Figure 2.

The main idea of AdapBridge is simple: we first use the model to generate all words with Equation 1, and then compute the pairwise cosine similarity between the generated words and the ground truth. If the generated word is learned good enough (similar to the ground truth or a synonym), the max cosine similarity of this word will be close to 1 and upper a threshold β showed in Figure 2. Therefore, we can use this generated word to replace the ground truth word, which introduces the generator distribution to the training phase. The summary of the algorithm is illustrated in Table 2.

3.2 Sampling with Increase

The threshold α and β determine the frequency of sampling function and the similarity between generated words with ground truth, respectively. Note that, when $\alpha = 0$, the training type is same as before, while when $\alpha = 1$, the model is trained as inference. If α is set too low(≈ 0), the inputs of decoder will almost be ground-truth, and will not be able to cope with the unknown words predicted in the reference. On the other hand, if α is set too high(≈ 1) at the beginning of training, the model will yield tokens randomly, because the model is not well trained, which may lead to slow convergence. Similarity, because model can not generate high cosine similarity score at the beginning, it is necessary to set the β low to ensure that part of ground truth words can be replaced by generated words, and increase its value as the training steps or epochs growing. In this sense, α and β should be both increase with the training time grows. Note that, the probability α in our method determine whether to execute the transition mechanism. And the the threshold β determine which words in the generated sentence should be replaced, with respect to the cosine similarity score.

Algorithm 1 AdapBridge

Input:

    Sequence of generated words $y_{<t}^{k}{}^{*}$;

    Sequence of ground-truth words $y_{<t}^{k}$;

    Word embedding of model with size of shared vocabulary; Number of epoch n.

Output:

    Inputs of decoder after sampled $y_{<t}^{k}{}^{\wedge}$

1:   Initialize $P \leftarrow \{p_1, p_2, \cdots, p_{|e_{<t}^{k}{}^{*}|}\}$

     Calculate α and β with n

     Get a random number m between [0,1].

2:   **if** $m < \alpha$ **then**

3:      $e_{<t}^{k}{}^{*} \leftarrow \mathbf{E}(y_{<t}^{k}{}^{*}), e_{<t}^{k} \leftarrow \mathbf{E}(y_{<t}^{k})$

4:      **for** $i = 1, \cdots, |e_{<t}^{k}{}^{*}|$ **do**

5:        **for** $j = 1, \cdots, |e_{<t}^{k}|$ **do**

6:          $\mathbf{SimMat}(i, j) \leftarrow \frac{e_i^{k*} \cdot e_j^{k}}{|e_i^{k*}||e_j^{k}|}$

7:        **end for**

8:        $s_i \leftarrow \underset{j \in 1, \cdots, |e_{<t}^{k}|}{\mathbf{MaxSim}}(\mathbf{SimMat}(i, j))$

9:        $p_i \leftarrow I_{>\beta}(s_i)$

10:     **end for**

11:     $y_{<t}^{k}{}^{\wedge} \leftarrow y_{<t}^{k}{}^{*} \otimes P + y_{<t}^{k} \otimes (1 - P)$

12: **else**

13:     $y_{<t}^{k}{}^{\wedge} \leftarrow y_{<t}^{k}$

14: **end if**

15: **return** $y_{<t}^{k}{}^{\wedge}$

Table 2: Algorithm of AdapBridge

3.2.1 Increase Function of α

We define α with an increase function dependent on the number of training epochs n:

$$\alpha = 1 - \frac{k}{k + exp((n - w)/k)} \qquad (5)$$

where $k \geq 1$ is hyper-parameter, which determines the speed of convergence. In addition, we add a parameter w, which makes α close to 0 during the first w epochs of training. It is usually set to half number of all training epochs to ensure that the model is trained enough to generate reasonable words. The curve of this function can be seen in Figure 3.

3.2.2 Increase Function of β

At the beginning of training, the words cosine similarity scores generated by the model are generally low, so reducing β appropriately could help. On the other hand, at the end of training, β should increase, as the model was already trained good
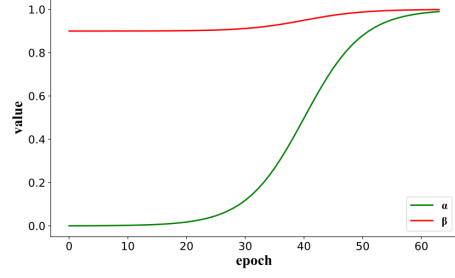
Figure 3: The curve of α and β. Where $k = 5, w = 32$ in Equation 5, and $\gamma = 0.9$ in Equation 6

enough to generate meaningful words after the first w epochs. However, β can not start from zero as α showed in Figure 3. Intuitively, even at the beginning, it should have a certain threshold to ensure that the generated words which will replace the ground truth words have a high quality.

We thus propose to use a schedule to increase β as a function of α calculated with Equation 5 , the formula as follows:

$$\beta = \gamma + (1 - \gamma) * \alpha \qquad (6)$$

where γ is the lowest similarity score threshold. In the entire training process, one ground truth word can be replaced only when the score of the generated word is at least greater than γ. The function is strictly monotone increasing the same as the Equation 5, and its curve can be seen in Figure 3.

4 Experiments

In this section, we evaluate our proposed model on both Chinese STC and the English Reddit dataset.

4.1 Experimental Settings

We first introduce datasets, baselines, parameters setting and evaluation measures.

4.1.1 Datasets

We use two public one-to-many single turn dialogue datasets in our experiments. The Chinese Weibo Dataset, named STC[1], consists 4,391,266 , 23,532 and 21,161 dialogue context-response pairs for training, validation and testing sets, respectively. We remove those pairs whose context contains response or contrary, and obtain 4,295,557 , 23,039 and 20,749 pairs for three data sets. The average number of responses corresponding to each context

---

[1]https://cloud.tsinghua.edu.cn/f/372be4a9994b4124810e/

in STC is 19.7. The English Reddit dialogue corpus is extracted from the Reddit comments-posts by the script[2], named Reddit. The original data consists of 6 million dialogues from all even months of 2011. We use the official script to tokenize, and the duplicates and sentences with length less than 3 or longer than 64 are removed. If the number of responses of one context is more than 20, we will randomly select 20 responses and ignore others, and the average number of responses corresponding to each context in Reddit is 6.2. Finally, we randomly split the data to training, validation and testing sets, which contains 1,107,860, 23,183, 12,429 pairs, respectively.

4.1.2 Baselines and Parameters Setting

Three baselines are used for comparison, including Transformer-based model (Vaswani et al., 2017), Random Sampling with word(RS-word) and sentence(RS-Sentence) (Zhang et al., 2019).

For STC, we utilize the Chinese word as input, and set the vocabulary size as 10,599. For Reddit, context-response pairs are encoded using byte-pair encoding(BPE) (Sennrich et al., 2016) with vocabularies of 11,527 tokens. For a fair comparison among all baseline models and our model, the dimension of all word embedding is 512, and beam size in testing is 5. The transformer model has 6 layers in both encoder and decoder, while 8 heads in multi-head attention. All parameters are initialized by the uniform distribution over$[-0.1, 0.1]$. We adopt the optimizer Adam (Kingma and Ba, 2015) with $\beta 1 = 0.9, \beta_2 = 0.98$ and with a weight decay of $\epsilon = 10^{-8}$. We set the learning rate as 0.0007 and the maximum tokens of a batch as 8192 with the update frequency 2. We run all models on 4 Tesla P40 GPU cards with PyTorch[3]. The code will be released when this paper is accepted.

4.1.3 Evaluation Measures

The evaluation of quantitative metrics and human judgements are used in our experiments. Specifically, quantitative metrics contains traditional metrics, such as PPL and BLEU score (Papineni et al., 2002), and Distinct (Li et al., 2016) metric which is recently proposed to evaluate the degree of diversity of the generated responses by calculating the number of distinct unigrams and bigrams in the generated responses. We also evaluate each generated response by calculating BLEU score with

all reference responses, and use the highest BLEU score to represent the quality of generated response. The average of all highest BLEU score in the testing set is named AH-BLEU. In addition, BLEU score is calculated by using the toolkit of NLTK[4].

For human evaluation, given 300 randomly sampled contexts and their responses which are generated by different models, three annotators (all CS majored students) are required to give the score of those context-response pairs, e.g. 3, 2, 1 means relevant, common and no-relevant, respectively, based on the coherence of the generated responses with respect to the contexts. The mean score is the average of all scores given by the three annotators with context-response pairs generated by a model. Meanwhile, in order to get the relative score of different models, we also evaluate the ground-truth context-response pairs by human evaluation.

4.2 Experimental Results

In this section, we demonstrate our experimental results on the two public datasets.

4.2.1 Metric-based Evaluation

Table 3 shows the quantitative evaluation results. From this table, we can see that models with switch mechanism, such as RS-Word, RS-Sentence and AdapBridge, outperform the traditional Transformer-based model in terms of BLEU, Distinct-1 and Distinct-2 evaluations. The results show that the switch mechanism plays an important role in the dialogue generation task.

RS-Word and RS-Sentence both replace the ground truth tokens by the generated tokens with a random scheduled sampling. However, their performances are both worse than our proposed model, as our model considers the relevance between the generated words and the ground truth with word-level matching scores. For the BLEU score on STC dataset as an example, the BLEU-4 score of Adap-Bridge is 2.17, which is better than that of RS-Word and RS-Sentence, i.e., 2.05 and 2.12. Especially, our model achieves best AH-BLEU-2 score on both two datasets, which is the significant performance gains. It shows that the responses of our model have higher quantity than other baselines.

The diversity of responses can be evaluated by Distinct score. As shows in Table 3, our Adap-Bridge achieves significant performance gains. Take the results of Reddit in Table 3 as an example, the proposed AdapBridge model improves the

STC Datasets						
Model	PPL	BLEU-2(%)	BLEU-4(%)	DIS-1(%)	DIS-2(%)	AH-BLEU-2(%)
Transformer	28.86	3.74	1.37	0.23	0.90	14.43
RS-Word Oracle	28.91	5.12	2.05	0.33	1.25	15.21
RS-Sentence Oracle	**26.75**	**5.50**	2.12	0.35	1.38	15.52
AdapBridge	29.36	5.35	**2.17**	**0.43**	**1.74**	**16.38**
Reddit Datasets						
Model	PPL	BLEU-2(%)	BLEU-4(%)	DIS-1(%)	DIS-2(%)	AH-BLEU-2(%)
Transformer	40.83	**3.99**	0.77	0.79	2.91	7.03
RS-Word Oracle	43.11	3.78	0.81	1.42	5.19	7.43
RS-Sentence Oracle	**40.72**	3.49	0.76	1.33	5.08	7.05
AdapBridge	48.01	3.56	**0.83**	**1.56**	**5.56**	**7.60**

Table 3: The metric based evaluation results on STC and Reddit datasets. DIS represent the distinct score and AH-BLEU-2 represent the average of all highest BLEU-2 score.

STC Datasets				
Model	3(%)	2(%)	1(%)	Mean
Ground Truth	82.23	13.56	4.23	2.78
Transformer	48.67	23.11	28.22	2.20
RS-Word	56.33	20.89	22.78	2.34
RS-Sentence	55.33	22.67	22.00	2.33
AdapBridge	59.56	24.33	16.11	**2.43**
Reddit Datasets				
Model	3(%)	2(%)	1(%)	Mean
Ground Truth	79.00	15.67	5.33	2.74
Transformer	49.78	21.89	28.33	2.21
RS-Word	52.44	23.00	24.56	2.28
RS-Sentence	53.11	24.67	22.22	2.31
AdapBridge	55.67	28.33	16.00	**2.40**

Table 4: The human evaluation on STC and Reddit.

Transformer, RS-Word and RS-Sentence models by 2.65, 0.37 and 0.48 Distinct-2 points, respectively. We can also note that our model has the highest Distinct score on both STC and Reddit datasets, which indicates that our model can generate more diverse response and avoid generating common responses. In summary, our proposed AdapBridge model has the ability to generate high quality and diverse responses, compared with baselines. We also conducted the significant test, and the result shows that the improvements of our model are significant on both two datatests, i.e., $p - value < 0.01$.

4.2.2 Human Evaluation

The human evaluation results are shown in Table 4. The percentages of relevant, common and no-relevant are given to evaluate the quality of responses generated by different models. From the results we can see that our AdapBridge gets the highest score in human evaluation. Take the STC as an example, compared with Transformer, RS-Word Oracle and RS-Sentence Oracle, the Adap-Bridge achieves performance gains 22.38%, 5.73%, 7.65% on the relevant score. For the Mean score, we can observe that our AdapBridge generates the most relevant responses, while generates less no-relevant responses, which indicates that the responses generated by our model are attractive to annotators. We also conducted the significant test, and the result shows that the improvements of our model are significant on both two datatests, i.e., $p - value < 0.01$.

4.3 Case study

In this section, we conduct case studies to show our model can generated more relevant and diverse responses than other baseline models.

We give two examples as in Table 5. In the Example 1, the response of Transformer is "Is this a fish or a fish?", which is an unreasonable sentence, as the words in the positions of two "fish" should be different according to common sense. For the response of RS-Word, we can see that it repeats "want to eat fish" twice, which is a part of the context. Although the response of RS-Sentence "How lovely! I want to eat, too." is relevant, it conforms

Example 1		
context	你不是天天嚷着要吃鱼嘛?鱼来了	Don't you want to eat fish every day? The fish is coming.
Transformer	这是鱼还是鱼啊?	Is this a fish or a fish?
RS-Word	要吃鱼，要吃鱼。	Want to eat fish, want to eat fish
RS-Sentence	好可爱！我也想吃。	How lovely! I want to eat, too.
AdapBridge	这是什么鱼?	What kind of fish is this?
Example 2		
context	水城威尼斯，梦里的城市。	Venice, the city of dreams.
Transformer	梦里的城市，梦里的城市。	City in dreams, city in dreams.
RS-Word	好美的城市，好想去。	Beautiful city. I really want to go.
RS-Sentence	这是哪呀！我也想去。	Where is this? I want to go too.
AdapBridge	想和我的爸妈一起去。	I want to go here with my parents.

Table 5: Two examples of generated responses on STC.

to a common response paradigm, such as "how · · · , I want it, too" or "what's this, I want to · · · ". If the context contains food, animal, locations etc., such responses all seem appropriate, which is not attractive to human. While the response generated by our AdapBridge is more specific and relevant, i.e. "What kind of fish is this?". We can also see the similar phenomenon in the Example 2 of Table 5, with the context "Venice, the city of water, the city of dreams.", Transformer repeats the same content that comes from the context, and the responses of RS-Word and RS-Sentence are both common responses as mentioned above. Compared with responses generated by baseline models, response of AdapBridge "I want to go here with my parents." is more relevant and attractive. Therefore, those results indicate that our proposed model can generate high quality and attractive responses with the adaptive switch mechanism.

4.4 AdapBridge on NMT

The method we propose can also be adapted for the neural machine translation(NMT) in an easy way. With this task we want to investigate if AdapBridge could be helping to improve the performance of NMT which is a classic natural language generation task. We perform experiments on the WMT'14 English→German(En→De) datasets, which contains 3,900,502, 39,414, 3,003 sentences for training, validation and testing sets, respectively. We train the Transformer-based model with the same setting described in Section 4.1.2, and then we measure the translation quality with BLEU. The evaluation results are listed in Table 6.

From the results, we can see that our method can also achieve significant performance gains, and im-

Model	BLEU-2(%)	BLEU-4(%)
Transformer	43.40	26.43
RS-Word	43.66	26.84
RS-Sentence	**44.08**	27.21
AdapBridge	43.99	**27.38**

Table 6: BLEU scores on En→De NMT task.

prove the Transformer-based model by 0.95 BLEU-4 points on average. For the BLEU-2 score, our model is slightly lower than RS-Sentences model same as the results in Table 3, it can attributed to the sentence-level information of RS-Sentence Oracles. In order to analyze the gap of ground-truth and the generated sentences, we calculate the cosine similarity between the hidden representations of ground truth sentences and generated sentences, with a trained Bert model (Wolf et al., 2020), and get the similarity score 0.96 and 0.81 on WMT'14 and Reddit datastes, respectively. At the same time, we can also notice that BLEU score of NMT is much higher than that of dialogue generation task. The results of overlap measures indicate the severity of the exposure bias problem in dialogue generation as analyzed in Section 1.

5 Conclusion

In this paper, we propose a novel adaptive switch mechanism with word-level matching scores to solve the problem of exposure bias for the dialogue generation task, named AdapBridge. Our core idea is to utilize the word-level matching scores to determine the input is from ground truth or from prediction at each step of training. Experimental results show that our model significantly outperforms pre-

vious baseline models. Further analysis on NMT also indicates that our model can achieve significant improvement on different generation tasks. In future work, we plan to further design different scoring methods, i.e. Bert score or BLEU, to guide the model selects better words. It is also interesting to extend our AdapBridge model to other generation tasks, such as abstractive summarization.

Acknowledgements

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1171–1179.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*, pages 2672–2680.

Sebastian Goodman, Nan Ding, and Radu Soricut. 2020. Teaforn: Teacher-forcing with n-grams. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8704–8717.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Weili Nie, Nina Narodytska, and Ankit Patel. 2019. RelGAN: Relational generative adversarial networks for text generation. In *International Conference on Learning Representations*.

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 of the Association for Computational Linguistics*, pages 311–318.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Coldgans: Taming language gans with cautious sampling strategies. In *Advances in Neural Information Processing Systems*, volume 33, pages 18978–18989. Curran Associates, Inc.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Chenze Shao, Xilin Chen, and Yang Feng. 2018. Greedy search with probabilistic n-gram matching for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4778–4784.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum

risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Arun Venkatraman, Martial Hebert, and J Bagnell. 2015. Improving multi-step prediction of learned time series models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'l'mi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lijun Wu, Yingce Xia, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Adversarial neural machine translation. In *Asian Conference on Machine Learning*, pages 534–549. PMLR.

Qingyang Wu, Lei Li, and Zhou Yu. 2021. Textgail: Generative adversarial imitation learning for text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14067–14075.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Self-adversarial learning with comparative discrimination for text generation. In *International Conference on Learning Representations*.