QUANTIFYING EXPOSURE BIAS FOR NEURAL LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The exposure bias problem refers to the training-testing discrepancy caused by teacher forcing in maximum likelihood estimation (MLE) training, for autoregressive neural network language models (LM). It has been regarded as a central problem for natural language generation (NLG) model training. Although a lot of algorithms have been proposed to avoid teacher forcing and therefore to alleviate exposure bias, there is little work showing how serious the exposure bias problem actually is. In this work, we first identify and analyze the self-recovery ability of MLE-trained LM, which casts doubt on the seriousness of exposure bias. We then develop a precise, quantifiable definition for exposure bias. Based on that definition, we design experiments to measure the seriousness of exposure bias. Surprisingly, we find that removing the training-testing discrepancy only brings very little performance gain, in both real and synthetic settings. With these results, we conclude that on the contrary to popular belief, the exposure bias problem is only a minor problem for MLE-based LM training.

1 INTRODUCTION

Language model (LM) is a central module for natural language generation (NLG) tasks (Young et al., 2017) such as machine translation (Wu et al., 2017), dialogue response generation (Li et al., 2017), image captioning (Lin et al., 2014), etc. For decades, maximum likelihood estimation (MLE) has been the most widely used objective for LM training. However, there is a popular belief in the natural language processing (NLP) community that standard MLE training will cause “exposure bias” and lead to a performance degradation during the test-time language generation.

The exposure bias problem (Bengio et al., 2015; Ranzato et al., 2016) refers to the following discrepancy between MLE training and test-time generation for language models: During training, the language model predicts the next word conditioned on history words sampled from the ground-truth data distribution. And during generation, the model generates words conditioned on history sequences generated by the model itself. However, due to the exposure to real data during training, the language model is biased to only perform well on the ground-truth history distribution. As a result, during generation the errors will accumulate along the generated sequence, and the distribution generated by the model will be distorted. The forced exposure to ground-truth data during training is also referred to as “teacher forcing”.

Given its definition, the exposure bias problem could rise in the general cases when the model needs to make a sequence of decisions or generations (e.g. music/pixel/speech generation (Lamb et al., 2016)). In this work, we focus on the task of language generation, because the exposure bias problem is originally proposed in this field (Bengio et al., 2015), and has since attracted huge research attention.

In order to avoid teacher forcing, many training algorithms (Bengio et al., 2015; Lamb et al., 2016; Ranzato et al., 2016; Yu et al., 2016; Zhu et al., 2018; Lu et al., 2018; Lin et al., 2017; Guo et al., 2017; Rajeswar et al., 2017; Wiseman & Rush, 2016; Nie et al., 2019; Shi et al., 2018) have been proposed as alternatives to MLE training. Most of these works utilize techniques from generative adversarial network (GAN) (Goodfellow et al., 2014) or reinforcement learning (RL) (Sutton & Barto, 1998). In this paper, we refer to these algorithms as non-MLE methods or text GANs.

Despite the huge research efforts devoted to alleviate exposure bias, surprisingly, its existence or significance is much less studied. In particular, to the best of our knowledge, no existing work
Table 1: Samples of a MLE-trained STOA transformer LM when fed with different types of length-10 history prefix. To save space, we omitted the first 7 words of the random history.

<table>
<thead>
<tr>
<th>Model Samples as History → Model Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>When asked about how she thinks about the games, Flocke dislikes most of those about it, citing instances of paranoia in her heart and trembling temper, which infuriated him.</td>
</tr>
<tr>
<td>Approximately 500 Finns became sick since early October when sleeping in their sleeping bags. On 3 October, the “Red Guard”, which had been organized two months previously by Marius Kuusinen ...</td>
</tr>
<tr>
<td>The entire key results of the arc be obtained through unifying methods to construct the prologue, three pieces could be combined instead of need to provide a final chapter.</td>
</tr>
<tr>
<td>Data Samples as History → Model Samples</td>
</tr>
<tr>
<td>Most of what is known of Du Fu's life is clear and graphic descriptions, memoirs, commentaries on storyboards, and descriptions of Canadian settlers. More than 60 biographies and ...</td>
</tr>
<tr>
<td>In the early 730s, he travelled in the Jiangsu province of Asia after Ashras ibn Abdallah al-Sulami visited Quanzhou in Bukhara, the capital of Turkmenistan and a native of the ...</td>
</tr>
<tr>
<td>Since the Song dynasty, critics have called Du Fu the “master poet”, a product of his use of Du Fu scenes to establish the empress’s nature and to emphasize his ...</td>
</tr>
<tr>
<td>Shuffled Data Samples as History → Model Samples</td>
</tr>
<tr>
<td>is Du ofFu of life what Most claimed was his tragic adaptation of John Ching’s The Janus of Hades, translated by disgraced performer just months before ...</td>
</tr>
<tr>
<td>in the he travelled, the early Jiangsu In 730s, he attended a mission on the peninsula. He soon moved to Monkwearmouth, on the northern shores of Baffin Bay in The ...</td>
</tr>
<tr>
<td>Since, the called have Song Fu critics dynasty Du Fu, who Zhang historians have included, have not rivaled HABS’s Web site held for 253 years. In 2015, HABS-based producers Oronoco ...</td>
</tr>
<tr>
<td>Random Sequences as History → Model Samples</td>
</tr>
<tr>
<td>...RANDOM... surface leader Game after a failed attempt to test her effectively in three fleets falling to I-30. This went unnoticed by most ichthyologists; none understood either strict rules ...</td>
</tr>
<tr>
<td>...RANDOM... faster elephant emperor decorations with Rocky Mountain state exploit by linking all black geese to 1970s planning regulations that prohibit slaughter of snake species. ...</td>
</tr>
<tr>
<td>...RANDOM... hitting remained prominently from the system as she witnessed no mention of criteria in the text. Douglas Turner noted then that Gottesfeld may have assumed ...</td>
</tr>
</tbody>
</table>

2 MOTIVATION: THE SELF-RECOVERY ABILITY

To study the seriousness of exposure bias in standard MLE LM training, we first stress that the following methodology, although tempting, is wrong: If we can rigorously show that the non-MLE methods proposed to avoid teacher forcing do indeed bring solid generation performance gain, then we can conclude exposure bias is a meaningful problem for the original MLE training. The reason is that we typically do not know the exact underlying reason for the performance gain. For example, despite the huge success of the batch normalization technique in deep learning, whether “internal covariate shift” (which is the motivation of batch norm) exists in deep neural network training remains a question (Santurkar et al., 2018). Therefore, in this work we seek a direct way to validate the seriousness of exposure bias.

We focus on the following informal claim that immediately follows from the original definition of exposure bias: During generation, if we set the history distribution to be the ground-truth data distribution instead of the model’s own distribution (now that there is no discrepancy between training and testing), then the model’s language generation quality should be much better (we will formalize this notion in Section 4 and 5).

We start with the following qualitative analysis. We feed a MLE-trained transformer LM on wiki-103 data-set (Baevski & Auli, 2018) with four kinds of prefixes: model’s own samples, data samples, shuffled (word-level) data samples or samples from a uniform random distribution. Then we let the model complete the sentence given these prefixes as history. We list some samples in Table 1 and more in Appendix A (this experiment is also repeated for a LSTM LM).
Assuming the seriousness of exposure bias, we expect the quality of generated sentence-completion samples with real-data prefixes to be significantly better than the ones from prefixes of model samples. However, by manual inspection, we do not observe noticeable differences in sample quality. More surprisingly, the model is still able to generate relevant and fairly high-quality samples from shuffled prefixes. Even in the extreme case where random sequences are fed, the model is able to generate reasonable sentences. Due to the recent increasing interest of solving exposure bias in the field of neural machine translation (NMT) (Zhang et al., 2019), we repeat the above experiment in a standard NMT setting in Appendix A, and get very similar observations.

These experiments clearly show that the MLE-trained auto-regressive LMs have the self-recovery ability, i.e. the model is able to recover from artificially distorted history input, and generate reasonably high-quality samples. This phenomenon is clearly in contradiction with the popular claim of exposure bias, that the error induced by the mismatch between history and data distribution should accumulate during the generation process.

Motivated by these experiments, in the following sections, we turn to more rigorous methods to quantify the significance of exposure bias. Note that our quantification approaches will be independent of the training procedure and only require inference from the trained model.

3 Notations

The task of auto-regressive language modelling is to learn the probability distribution of the \((l+1)\)th word \(W_{l+1}\) in a sentence conditioned on the word history \(W_{1:l} := (W_1, \ldots, W_l)\). Here, we use the uppercase \(W_i \in V\) to denote a discrete random variable distributed across the vocabulary \(V\). The lower-case \(w\) is used to denote some particular word in \(V\). Given a training data-set \(D\) consisting of sentences of length \(L\), the standard MLE training minimizes the negative log-likelihood below:

\[
L_{\text{MLE}} = \mathbb{E}_{W_{1:L} \sim D} \frac{1}{L} \sum_{l=1}^{L} \log P_M(W_l | W_{1:l-1}),
\]

(1)

Note that in this work we assume all sentences are of length \(L\) for simplicity.

We denote the generation distribution of the trained LM as \(P_M\), and the ground-truth data distribution as \(P_D\). Readers can assume \(P_M\) refers to the generation distribution of a LSTM LM (Hochreiter & Schmidhuber, 1997; Sundermeyer et al., 2012) or a transformer LM (Baevski & Auli, 2018; Dai et al., 2019) trained with MLE objective, which is the major subject of this study. We will mainly present results on LSTM based models to facilitate comparison with text-GAN works (listed in Section 1), which are mostly implemented on LSTM models. We will also provide results with the transformer model, with very similar observations or measurements.

Our quantification mainly relies on the measurements of the distance from the model’s generation distribution to the data distribution. Hence we define the following notations to simplify expressions. Let \(\mathcal{P}\) denote the set of probability distributions on the vocabulary \(V\). Let \(d\) denote a distance measure between distributions (e.g. total variation distance), \(d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}\).

4 A (Wrong) Quantification Using Marginal Distribution

In this section, we propose an intuitive and seemingly correct quantification approach using marginal distributions. The approach can be applied to real-world text data experiments, but it has some lethal weak point. The discussion will lead us to our final precise definition of exposure bias in Section 5.

4.1 Method

Assuming a given history length \(l\), we consider the marginal distribution of \(W_{l+1}\) from the following three random process:

- Draw word sequences of length \(L\) from the data distribution \(P_D\). Denote the marginal distribution of the random variable at position \(l+1\) \((W_{l+1})\) as \(P_{D|D}^{l+1}\), where

\[
P_{D|D}^{l+1}(w) = \mathbb{E}_{W_{1:l} \sim P_D}[P_D(w | W_{1:l})],
\]

(2)

- Draw word sequences of length \(L\) from the model distribution \(P_M\). Denote the marginal distribution of the random variable at position \(l+1\) as \(P_{M|M}^{l+1}\), where

\[
P_{M|M}^{l+1}(w) = \mathbb{E}_{W_{1:l} \sim P_M}[P_M(w | W_{1:l})].
\]

(3)
First draw $W_{1:l}$ from $P_D$, then draw $W_{l+1}$ from $P_M(W_{1:l})$. Denote the marginal distribution of the random variable at position $l + 1$ as $P^{l+1}_{M|D}$, where
\[
P^{l+1}_{M|D}(w) = \mathbb{E}_{W_{1:l} \sim P_D}[P_M(w | W_{1:l})].
\] (4)

By the definition of exposure bias, $P^{l+1}_{M|M}$ suffers from the training-testing discrepancy, while $P^{l+1}_{M|D}$ should be closer to the true distribution $P^{l+1}_{D|D}$. To measure this discrepancy, define the marginal generation deviation (MGD) at history length $l$ of history distribution $P_H$ with metric $d$ as
\[
\text{MGD}(P_M[H,l], d) = d(P^{l+1}_{M[H]}, P^{l+1}_{D[D]})
\] (5)

where $P_H \in \{P_M, P_D\}$ denotes the history distribution. MGD measures the deviation of the marginal distribution of $W^{l+1}$ from ground-truth data distribution.

Finally, we define the rate of exposure bias (EB-M) at history length $l$ of model $P_M$ as the ratio (discrepancy) between the MGD measurements when two different history distributions are fed:
\[
\text{EB-M}(P_M, l, d) = \frac{\text{MGD}(P_M[H,l], d)}{\text{MGD}(P_M[D,l], d)}
\] (6)

For MLE-trained models, EB-M \(^1\) is expected to be larger than 1, and larger EB-M indicates a more serious exposure bias problem for the trained model. For the metric $d$, we consider two popular probability metrics: total variation distance (denoted as $d_{TV}$), and Jensen-Shannon divergence (denoted as $d_{JS}$).

### 4.2 Experiments and Discussion

In this section, we focus on answering the following question: “Does the EB-M measurement correctly reflect the significance of exposure bias?” In short, our answer is not really. The problem is that the distortion of the marginal $P^{l+1}_{M|M}$ is not only affected by the presumably existing exposure bias problem alone, but also by the mismatch between the history distribution $P_M$ from $P_D$ for $W_{1:l}$, which grows with the length of the history. Therefore, even if the measured EB-M is significantly larger than one, we cannot conclude that exposure bias causes serious deterioration. We provide an example to illustrate this argument:

**Example 1.** Suppose $L = 2$, and $V = \{A, B\}$. $P_D$ and $P_M$ are crafted as follows: $P_D(AA) = P_D(AB) = 0.5$, $P_D(AB) = 0$; and $P_M(W_1 = A) = 1$, $P_M(W_2 = A | W_1 = A) = 1$, $P_M(W_2 = B | W_1 = B) = 1$.

In Example 1, MGD($P_M[M,1], d_{TV}$) = 0.5 and MGD($P_M[D,1], d_{TV}$) = 0, which gives EB-M($P_M, 1, d_{TV}$) = $\infty$. However, the only problem $P_M$ has is the mismatch between the history distributions ($P_M$ and $P_D$) for $W_1$.

The next set of experiments also suggest that EB-M does not precisely reflect exposure bias. On the EMNLP-news data-set (specified in Appendix B), we compare EB-M measurements for several non-MLE training methods with the baseline MLE model. We include results for Scheduled Sampling (SS) (Bengio et al., 2015), Cooperative Training (CoT) (Lu et al., 2018), and Adversarial Ranking (RankGAN) (Lin et al., 2017). We provide implementation details for non-MLE methods in Appendix C. Intuitively, these methods will cause the model to be biased to behave well with model samples as history, instead of data samples. Therefore, we expect EB-M measurement for non-MLE trained models to be smaller than MLE trained models. However, Figure 1 shows that the measurements for different training frameworks are almost the same. We believe the reason is that the EB-M measurements are only reflecting the trivial mismatch between the history distributions.

Is it possible that the original definition of exposure bias (Bengio et al., 2015; Ranzato et al., 2016) exactly refers to this mismatch between the model and data history distributions? However, note that this mismatch is inevitable for any imperfect model, and non-MLE training algorithms can not solve it. We believe a better, more precise definition is needed to discriminate exposure bias from this trivial mismatch. Motivated by this view, we propose a second approach in the section below.

\(^1\)Note that one can also directly measure $d(P^{l+1}_{M|M}, P^{l+1}_{M|D})$, but in that way, we can not tell which distribution is better.
5 A Quantification Approach Using Conditional Distribution

5.1 Method

Following the discussion in the last section, we wish our measurement to be independent of the quality of the history distribution. In light of that, we design a quantity to measure the model’s conditional generation quality. Let $P_H \in \{P_M, P_D\}$ denote the history distribution as in the MGD definition (5). With history length $l$ fixed, we define the conditional generation deviation (CGD) with history distribution $P_H$ for $P_M$ using metric $d$ as:

$$\text{CGD}(P_M|H, l, d) = \mathbb{E}_{W_{1:l} \sim P_H} [d(P_M(\cdot|W_{1:l}), P_D(\cdot|W_{1:l}))]$$

where we assume that $P_D(\cdot|W_{1:l})$ is computable, and use it to measure the quality of the model’s conditional distribution. For the choice of the distribution distance $d$, in addition to $d_{TV}$ and $d_{JS}$, we introduce greedy decoding divergence ($d_{GD}$) defined as:

$$d_{GD}(P, Q) = 1(\arg \max_i P_i \neq \arg \max_i Q_i)$$

where $1$ is the indicator function, and $P, Q \in \mathcal{P}$. The distance $d_{GD}$ reflects the model’s accuracy during greedy decoding.

Similar to MGD, exposure bias should imply a significant gap between CGD($P_M|M, l, d$) and CGD($P_M|D, l, d$). We again define rate of exposure bias at history length $l$ with metric $d$ to be:

$$\text{EB-C}(P_M, l, d) = \frac{\text{CGD}(P_M|M, l, d)}{\text{CGD}(P_M|D, l, d)}$$

For our definition of EB-C, a natural question is why we only focus on the generation distribution of the very next word. The reason is we want to precisely measure how the error caused by the history part affect the generation part, by keeping them separate. If we measure the deviation of, for example, two sampled tokens, the definition will be confusing: Because the second sampled token will be affected not only by the accumulated error induced by the history (sampled from the model), but also by the first generated token as history. To get a better understanding of the intuition behind the definition of EB-C, we recommend readers to read Appendix A about our NMT experiment.

5.2 Synthetic Experiments and Discussion

Since CGD requires inference for ground-truth data distribution $P_D$, we first consider experiments in a synthetic setting. In text-GAN literature (Yu et al., 2016; Lin et al., 2017), a randomly-initialized
one-layer LSTM model with hidden dimension of 32 is usually used as $P_D$ in synthetic experiments (we denote this setting as $M_{32}^{\text{random}}$). However, the model is small-scale and does not reflect any structure existing in real-world text. To improve upon this approach, we take the MLE baseline model trained on EMNLP-news data (described in Appendix B) as $P_D$ in this synthetic setting. We denote the data model ($P_D$) as $M_{\text{news}}$. We then train two LSTM LM ($P_M$) with different capacities using samples from the data model, with the standard MLE objective. One is a one-layer LSTM with hidden width of 512 (denoted as LSTM-512), the other one is with hidden width of 32 (denoted as LSTM-32).

We train $P_M$ for 100 epochs using the Adam optimizer with learning rate 0.001. In each epoch, 250k sentences (same to the size of the original EMNLP-news data) of length $L = 50$ are sampled from $M_{\text{news}}$ as training data to avoid over-fitting. We show perplexity (PPL) results of the trained models in Appendix F. Finally, EB-C is calculated using 100k$^3$ samples from $P_M$ and $P_D$.

![Figure 2: EB-C measurement for the LSTM-32 and LSTM-512 model with different metrics. Also average value of EB-C along all history length is shown in the legend.](image)

In Figure 2, we show EB-C measurements with different metrics $d, m$, and the two models give similar results. It is shown that EB-C has a steady but slow increasing trend as history length increases. However, the average value of EB-C is less than 1.03 (the largest average value is from $d_{JS}$ for the LSTM-512 experiment), meaning that the gap between CGD($P_M|M, \ell, d$) and CGD($P_M|D, \ell, d$) is not large. Also, note that in most NLG applications (such as machine translation or image captioning), the generated sequence typically has short length (less than 20). In that range of history length, the EB-C measurements that exposure bias only has minimal influence.

In Appendix E, we repeat the experiment for a transformer LM (Dai et al., 2019), and get very similar EB-C measurements. These measurements imply a striking conclusion: (Informal) Even if all the bad effects from exposure bias for MLE LM training are removed, the relative performance gain is at most 3%. If the sequence length is not very long, the gain is less than 1%.

To dive deeper into the cause of the gap in CGD, we experiment with corrupted versions of $P_M$ as history distribution. We first specify a corrupt rate $c \in [0, 1]$, and randomly substitute words in a history sample from $P_M$ to a “noise” word drawn uniformly from the vocabulary with probability $c$. Consequently, larger $c$ will cause the history distribution to deviate farther from the ground-truth $P_D$. In Figure 3, we show CGD measurement versus the corrupted history $P_M^{\text{corrupt}}$. Large gaps are observed between CGD($P_M|M, \ell, d$) and CGD($P_M|D$). Therefore, the small gap between CGD($P_M|M$) and CGD($P_M|D$) in Figure 2 results from the small deviation between the history distribution $P_M$ and $P_D$. In other word, $P_M$ has learned a “good enough” distribution that is able to keep it in the well-behaving region during sampling.

With these observations, we conclude that, in the synthetic setting considered, exposure bias does exist, but is much less serious than it is presumed to be. Although there exists mismatch between the history distribution $P_M$ and $P_D$, the mismatch is still in the model’s “comfortable zone”. In

\[ \text{We show that we can get stable measurements using 100k samples in Appendix E.} \]
other words, the LSTM LM is more robust than exposure bias claims it to be. To concretize this argument, we provide an example LM and show that MLE training is unlikely to generate models with a large EB-C value.

**Example 2.** Again suppose $L = 2$, and $V = \{A, B\}$, the ground-truth data distribution is uniform on $\{AA, AB, BB, BA\}$. $P_M$ is crafted as follows: $P_M(W_1 = A) = 0.9$, $P_M(W_2 = A|W_1 = A) = 0.9$, $P_M(W_2 = A|W_1 = B) = 0.5$. Note that the model behaves bad when $W_1 = A$, which is of high probability during sampling.

In **Example 2**, $\text{CGD}(P_M|D, 1, d_{\text{TV}}) = 0.2$ and $\text{CGD}(P_M|M, 1, d_{\text{TV}}) = 0.36$, so $\text{EB-C}(P_M, 1, d_{\text{TV}}) = 1.8$. However, this crafted model is unlikely to be an outcome of MLE training. The fact that $P_M(W_1 = B)$ is better modeled indicates that in the training data more sentences begin with $W_1 = B$ than $W_1 = A$. So MLE training should assign more probability to $P_M(W_1 = B)$, not the other way around.\(^5\)

Finally, we use EB-C to compare MLE and non-MLE training. We compare MLE against CoT, SS, RankGAN in the synthetic experiments, and results are shown in Figure 4. Note that the RankGAN experiments are conducted in the $M_{\text{random}}$ setting, as we find it hard to do a fast implementation of RankGAN for the LSTM-512 setting. We find that RankGAN and CoT gives lower EB-C measurements than MLE, which is expected, as these methods avoid teacher forcing. For CoT, at short

\[\text{If we change to } P_M(W_1 = A) = 0.1,\text{ then } \text{EB-C}(P_M, 1, d_{\text{TV}}) \text{ will be 0.2, meaning that the model has better conditional generation performance during sampling.}\]

\[\text{The MLE model is used as the pre-trained model for the RankGAN generator. The MLE model has an oracle NLL of 8.67, and RankGAN’s oracle NLL is 8.55.}\]
History: in august 2009, vh1 classic aired music videos from the tv special around the beatles (1964).
Choices: 0: the 1: including 2: which 3: featuring 4: and 5:[None of the above is plausible]

History: contingent units were accused of being armed with antiquated weapons, while some local peasants took claims of having been
Choices: 0: conscripted 1: armed 2: involved 3: tricked 4: intimidated 5:[None of the above is plausible]

Table 2: An illustration for the next word collection process. The choices are shuffled. The first history sample is from real data, and the second history sample is from the trained model.

| History Length (l) | CGD($P_{M|D}, l, d_{GD}$) | CGD($P_{M|M}, l, d_{GD}$) | EB-C |
|-------------------|--------------------------|--------------------------|------|
| 10                | 0.723                    | 0.738                    | 1.019|
| 20                | 0.736                    | 0.752                    | 1.021|
| 30                | 0.732                    | 0.754                    | 1.029|

Table 3: EB-C measurements with human as $P_D$.

hisotory length, EB-C is even less than 1. We believe the reason is that CoT trys to make the model be biased to behave better when fed with model samples. However, SS gives worse EB-C measurements comparing to MLE, which we currently do not have a good explanation. We refer readers to Huszr (2015) for a discussion about the SS objective.

To the best of our knowledge, this is the first direct empirical evidence that text GAN does indeed alleviate the exposure bias problem. It also indicates that EB-C correctly reflect the significance of exposure bias. We believe the reason for why EB-C is still not less than 1 is that, text GANs still rely on MLE pre-training a lot.

5.3 EB-C Measurements with Real Human as $P_D$

In this section, we design experiments to efficiently estimate EB-C for a SOTA transformer LM with real human as $P_D$, by utilizing the Amazon Mechanical Turk (AMT) platform. Given a MLE-trained LM as $P_M$, by examining the definition of EB-C in Equation 9 and 7, it is clear the only obstacle is that we don’t have access to $P_D(\cdot | W_{1:l})$ with a given history $W_{1:l}$. So, in this section, we focus on the greedy decoding divergence ($d_{GD}$) metric (Equation 8), which only requires the turkers to give the most probable next word prediction, instead of the full distribution (which is clearly intractable).

In our preliminary trials, we find it is still very hard for a person to guess the next word, even with real data history samples. The reason is that the vocabulary is very big, and the turkers may be not familiar with the context (e.g. wikipedia). To alleviate that problem, we design the following simplification: For a given history, we let the model output its top-5 next word prediction, then we only ask the turkers to choose among the 5 choices (the turker can also express that he/she thinks none of them is likely). Finally, we examine whether the turker’s choice is indeed the model’s top-1 prediction. We illustrate this process in Table 2.

We use the code of Transformer-XL (Dai et al., 2019) to train a SOTA transformer LM on the wiki-103 data-set. We favour the wiki-103 data-set because it is large-scale and has long (over 30 words) paragraphs, which is useful for the measurements of exposure bias. The model is a 16-layer transformer-xl model with hidden dimension of 410. Since the estimation of CGD($P_{M|D}, l, d$) requires large amounts of unseen real data samples, we use half of the wiki-103 training data (around 900k sentences and 50m words) to train the model $P_M$, and save the other half as samples from $P_D$. Other training configurations (learning rate, batch size, etc.) are not changed. The resulting model $P_M$ has a test-set PPL of 27.81 (if trained on full training data, the PPL will be 24.02).

We collect data to estimate EB-C at history length 10, 20, and 30. For each length and history model ($P_M$ or $P_D$) pair, we collect 10k $d_{GD}$ samples (via next-word prediction) from turkers on the AMT platform. More details about the AMT setup are provided in Appendix D. The results are shown in Table 3. The EB-C measurements are strikingly similar to the results in our synthetic experiments in that, removing the training-testing discrepancy only gives around 2% of relative performance gain.

[https://github.com/kimiyoung/transformer-xl/blob/master/pytorch/run_wt103_base.sh]
This result further strengthens our claim that exposure bias is only a minor problem for MLE-based LM training.

6 RELATED WORKS

Several recent works attempt to carefully evaluate whether the non-MLE training methods (e.g. adversarial training) can give superior NLG performance than standard MLE training for RNN LM. Caccia et al. (2018) tunes a “temperature” parameter in the softmax output, and evaluate models over the whole quality-diversity spectrum. Semeniuta et al. (2018) proposes to use “Reverse Language Model score” or “Frechet InferSent Distance” to evaluate the model’s generation performance. Tevet et al. (2018) proposes a method for approximating a distribution over tokens from a GAN, and then evaluate the model with standard LM metrics.

These works arrive at a similar conclusion: The general performance of Text GANs is not convincingly better, or even worse, than standard MLE training. Hence to some extent, they imply that exposure bias may be not a serious problem in MLE training. However, as we argued in Section 2, one can not draw direct conclusions about exposure bias with these results. For example, it is also possible that exposure bias is indeed serious for MLE training, but text GAN does not solve the problem well enough.

7 CONCLUSION AND DISCUSSION

In this work, we first identify the self-recovery ability of MLE-trained LM, which casts doubt on the seriousness of exposure bias, which has been regarded as a central problem for MLE training by the LM community. We then explore two intuitive approaches to quantify the significance of exposure bias for LM training. The first quantification EB-M relies on the marginal generation distribution and reveals some vagueness in the original definition of exposure bias. We argue that we should focus on the model’s generation performance in terms of its conditional distribution and propose a second quantification EB-C, which we regard as the precise definition for exposure bias.

We design a evaluation of EB-C at different history length with real human (turkers from AMT) as the data model, for a SOTA transformer LM. It is shown that removing the training-testing discrepancy only gives around 2% of performance gain. Our synthetic experiments also gives very similar measurements. By analyzing EB-C measurements with perturbed history samples, we hypothesise that although the mismatch between the data and model distribution for history prefix exists, it is still in the model’s “comfortable zone”. With these results, we claim that on the contrary to the popular belief, exposure bias is only a minor problem in MLE-based LM training.

To wrap up, we discuss the fundamental question “Is MLE training really biased?”, from the perspective of objective functions. Note that the MLE objective (1) can be re-written as:

\[
\begin{align*}
\arg\min_{\theta} \mathbb{E}_{W_{1:L-1} \sim P_D} \log P_M(W_{1:L-1}) & = \arg\min_{\theta} \mathbb{E}_{W \sim P_D} - \log P_M(W) \\
& = \arg\min_{\theta} \mathbb{E}_{W \sim P_D} \log \frac{P_D(W)}{P_M(W)} = \arg\min_{\theta} D_{KL}(P_D||P_M)
\end{align*}
\]

where \( D_{KL} \) denotes the Kullback-Leibler divergence, and \( \theta \) denotes the trainable parameters in \( P_M \). Therefore, MLE training is minimizing the divergence from \( P_M \), which is exactly the model’s sampling distribution, from \( P_D \). While it’s true that the training is “exposed” to data samples, we can not simply deduce the objective is “biased”.

We want to end our discussion with two remarks. First, the proposed quantification approaches should not be used as the only metric for NLG. For example, a position-aware uni-gram LM, which generates words independent of previous context, has no exposure bias problem and can pass our test easily. Second, the intention of this work is not to discourage researchers from exploring non-MLE training algorithms for LM. It is completely possible that an training objective different from \( D_{KL}(P_D||P_M) \), such as \( JSD(P_D||P_M) \), can lead to better generation performance (Lu et al., 2018; Huszr, 2015). However, though non-MLE algorithms avoid teacher forcing, these algorithms (using GAN or RL for example) are usually less stable and more difficult to tune. Given that the quantified measurement of exposure bias is insignificant, we think it should be questioned whether adopting these techniques to avoid exposure bias is a wise trade-off.
REFERENCES


### A The Auto-recovery Ability in General LM and NMT

In Table 6, we provide more samples of a MLE-trained transformer LM model (discussed in Section 2) when fed with different kinds of history. And in Table 7 we repeat the experiment for a LSTM-LM trained on the EMNLP-News data.

In Table 4 we repeat the preliminary experiment in Section 2 for a standard NMT setting. We train a 6-layer transformer model with hidden dimension 1024 on the IWSLT’14 German to English data set. We feed the trained model with types of prefix during decoding which represents different level of training-decoding discrepancy. Note that the source input is kept intact.

The result is very similar (or more striking) to our language model experiment, the data prefix does not seem to help, and in the extreme case of random prefix, the model still generates fairly good translation. In Section 2 we summarize this observation as the auto-recovery ability.

To interpret the UNREL3 results, we should not directly compare the translation generated from unrelated prefix to the target translation. In fact, we cannot even compare part of it (e.g. the part after the length-3 prefix). Instead, we highlight the surprising fact that although the model is forced to begin (conditioned) with a wrong prefix, it still comes up with a reasonable translation. This is not an
Table 4: A standard NMT transformer model fed with different types of length-3 history prefix. We did not do any cherry picking. The “@@” is because BPE tokenization is used. “DATA” means the first three output tokens are forced to be correct. “NORMAL” means no prefix is forced during decoding. “UNREL” means the first three tokens are forced to be from another random unrelated sentence (which is wrong but grammatical). “RAND” means the first three tokens are completely random words.

Table: easy task even for human translators, yet the model does fairly well. Again, this contradicts with the “exposure bias” hypothesis that a MLE-trained LM will produce a increasingly deviated sequence when initiated with a non-perfect prefix. Actually, during generation the model self-corrects the error in the prefix. It is also the major motivation of our proposed EB-C measurement (Section 5), which is based on the view of measuring distances between conditional distributions.

B Real-data Experiments for EB-M

One problem in the implementation of EB-M is to estimate the described marginal distributions of $W_{l+1}$. We adopt a simple sample-and-count method: $P_{D|M}^{l+1}$ is estimated by the distribution (histogram) of $W_{l+1}$ from a number (to be specified below) of sentences sampled from the data distribution. For $P_{D|M}^{l+1}$ and $P_{D|M}^{l+1}$, we first draw a number of history samples $W_{1:l}$ from the corresponding history model (model distribution and data distribution respectively). We then feed sampled history sequences into the trained model and estimate the marginal distribution of the $(l+1)$th word by averaging the predicted distribution $P_{D}^{l+1}(·|W_{1:l})$.

We measure EB-M for MLE-trained LSTM LM on two popular data-sets: EMNLP-news (EMNLP 2017 WMT News Section), and wikitext-103\(^7\). For EMNLP-news we set $L = 20$, and only use data samples whose length is longer than $L$. The resulting training/validation/test set has 268k/10k/10k

\(^7\)The wikitext-103 data is available at https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/.
sentences. The vocabulary is of size 5k. We use the 10k samples in the test set for evaluation of EB-M. Note that the EMNLP-news data-set is widely used in text GAN literatures Yu et al. (2016); Lu et al. (2018). We train a one-layer LSTM LM (Sundermeyer et al., 2012) of hidden dimension 512 as the MLE baseline model for EMNLP-news.

For wikitext-103, we set $L = 50$, and regard a paragraph in the original data as a long sentence. Further, we use half of the data for LM training, and utilize the other half for EB-M evaluation. The resulting training/validation/test/evaluation set has 300k/1.5k/1.5k/300k sentences. The vocabulary is of size 50k. We train a two-layer LSTM LM of hidden dimension 1024 as the MLE baseline model for wikitext-103.

For MLE baseline model training, the Adam optimizer is used with learning rate 0.001, no Dropout (Srivastava et al., 2014) is applied. The model is trained for 100 epochs.

We first measure EB-M on the wikitext-103 data-set, which has large amount of evaluation data. The results are shown in Figure 5. We provide EB-M measurements with metric $d_{TV}$ in Appendix E, as they are similar to those using metric $d_{JS}$. It is shown that the measurements become stable when using 100k data/model samples. EB-M has an average value of 1.10, indicating a significant gap between the model’s MGD when fed with history from $P_{D}$ or $P_{M}$. Further, we observe a steady growth of EB-M along the length of history, which is expected as an outcome of exposure bias.

![Figure 5: EB-M measurements (with metric $d_{JS}$) using different number of samples on wikitext-103 data.](image)

However, as discussed in Section 4.2, these measurements can only show that the LM does have better (marginal) generation quality when fed with data prefixes, but does not provide informative information for the significance of exposure bias.

C IMPLEMENTATION OF SS, CO-T, AND RANKGAN

We implement our MLE baseline and scheduled sampling (SS) in PyTorch. For SS, we use a linear decay schedule to move from complete teacher forcing to replace-sample rate of 0.1. We find that larger rate will give worse performance.

For CoT, we use a PyTorch implementation in https://github.com/pclucas14/GansFallingShort. We use a mediator model that has twice the size of the generator. We set M-step to be 4, and G-step to be 1.

For RankGAN, we use a TensorFlow implementation in https://github.com/desire2020/RankGAN.

Note that in our non-MLE experiments, the generator model is set to be the same size with the baseline MLE model. We tune the non-MLE methods using the corpus-BLEU metric, which is widely used in text GAN literature.
In this section we provide more details for the AMT evaluation discussed in Section 5.3. We show the HIT interface in Figure 6. Each HIT will include 10 pairs of context and its corresponding choices. Five of them are history samples from real data, and the other five is from the trained model. The history samples are mixed, so that the turker doesn’t know whether the history sample is from real data or the model. The next-word choices are also shuffled. The history length of the context could be 10, 20, or 30.

We collect around 10k HITs for each history length configuration. The same history sample is not repeated across the HITs. We limit each turker to do at most 200 HITs. For all history length configurations, there are around 300 unique turkers. As shown by Figure 7, most turkers conduct less than 20 HITs.

In Figure 8, we show that we are able to get stable measurements of EB-C with 100k samples for the LSTM-512 synthetic experiment.

In Figure 9 and Figure 10 we provide EB-M measurements with metric $d_{TV}$ discussed in Section 4.2, the results are similar to those using metric $d_{JS}$.

In Figure 11, we provide EB-C measurements of a 3-layer transformer LM (Dai et al., 2019) with 512 hidden dimension, in the synthetic setting.
Figure 7: A histogram of the number of HITs done by a unique turker.

Figure 8: EB-C measurements with different number of samples for the LSTM-512 synthetic experiment.

F Perplexity of the Trained Models

We show PPL results for model trained on EMNLP-news data-set in Table 5. The MLE model for wiki-103 data-set discussed in Section 4.2 has PPL 84.58. Note that due to our special setting\(^8\), our PPL result is not directly comparable to state-of-art LM results on these data-sets.

\(^8\)We only keep sentences of length longer than \(L\), and for wiki-103, only half of training data is used.

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE Baseline (M_{\text{News-512}})</td>
<td>55.85</td>
</tr>
<tr>
<td>LSTM-512 (MLE, synthetic)</td>
<td>115.3</td>
</tr>
<tr>
<td>LSTM-32 (MLE, synthetic)</td>
<td>156.3</td>
</tr>
<tr>
<td>Transformer-512 (MLE, synthetic)</td>
<td>103.6</td>
</tr>
<tr>
<td>CoT-512 (synthetic)</td>
<td>115.6</td>
</tr>
<tr>
<td>SS-512 (synthetic)</td>
<td>113.7</td>
</tr>
<tr>
<td>CoT</td>
<td>56.83</td>
</tr>
<tr>
<td>RankGAN</td>
<td>53.43</td>
</tr>
<tr>
<td>SS</td>
<td>56.43</td>
</tr>
</tbody>
</table>

Table 5: PPL results for model trained on EMNLP-news data-set.
He had a recurring role in 2003 on two episodes of the sitcom Roseanne as James "Bitch" Cook and guest starred in the 1999 special Richard Ayoade's comedy Canal &... 

Du Fu's compassion, for himself and for others, which arrived at Du Fu soon after the collapse of his political system. He was and still is called "<unk>"...

Du Fu's work is notable above all for its use of the convention people as arbiters in decision-making.

The tenor of his work changed as he developed his falsetto; the same February Bach performed his entire Magnificat in the Domus Aurea, a musical hall in Campo Bartolommeo, a ...

Although he wrote in all poetic forms, Du Fu explains that he had no intentions of writing poetry, and attempted to cash in on the success of his two-volume translation...

About two thirds of Du Fu's 1500 extant works survive as collections, but about one third have been rebuilt or linked. Some miniatures, such as the Memorial by...

According to the Encyclopaedia Britannica, Du Fu's writings joined a theme of opposition to social systems on the basis that the United States lacked standards and cotton-beets were...

Model Samples as History → Model Samples

Competing at the 2006 Commonwealth Games, McBreen scored 3.87 goals per game, ranking fourth from the conference in scoring, beaten 6:3 by Scotland. He also...

He, along with some young Christians from Poland, Romania, and East Germany, were taught to play dilruba. In order to achieve this, the boy recorded 40 or 50 dilruba parts...

EEF service throughout this filter was to suffer. This was approximately with the British renewal and capture charges on Mount Ciss, which contributed a large strength taking time to fall...

The matriarchal nature of the family is tested as opposed to that of their neighbors. In-laws explain their position by having the rear bedroom bathed in bondage to reflect cosmic...

The branch office distributed tuition to the top level schools, gaining coverage in the art of instruction in schools which allow them to select classes exclusively on the basis of...

Shuffled Data Samples as History → Model Samples

of Below an one of example is Du Fu's Systme, also the address of the No. 1 monuments Society and Advertising identifies a mass scale summarises his He by that Hung concluding life, let alone die. He ends by dying as saying "I died on the way". Robert Penson has selected Hung's last words...

Ann to designed music accompanying group the, video display was full on bluescreen and was rendered with Ghibli HDTVs. For the Xbox 360's Steam control, the cloud density was increased....

Random Sequences as History → Model Samples

...RANDOM... execution love Author → Churches Under Sunset and Angels post the 20 @,000 To 30 @,000 Arc landscape-crosses around 500 Enix areas...

...RANDOM... beyond spiders annually as part of regional zoning plans, including a pie canning pool in Mechanicsburg, some Ellisburg, and boxes of all medical...

...RANDOM... realm unknown healthy-bred Spock (released in 1991 as The Return), arrives in Sickbay to find a team; he engages in normal conversation (the main...

...RANDOM... rough elections appointment levels as he had already secured the if no candidate received the season ticket, a result of the September 11 attacks. ...

...RANDOM... / horses Finn s experience of sexual frog foraging, and might pose a threat to sexual preference as the crop Betsimisaraka earn "...

...RANDOM... Poland 1963 medium. → Basu was the visual effects supervisor on 300 visual effects shots of Gangster, Feller's seventh appearance in a Bollywood film. Aamir...

...RANDOM... levels MD defending her city of Beaufort, East Carolina in 2004. At the same time, she responded against the package of short-form compatible boats...

Table 6: More samples of a STOA MLE-trained transformer LM (on the wiki-103 data-set) when fed with different kinds of history. To save space, we omitted the first 7 words of the random history.
it was only a pieces that had gone up to the forest and forces the shoppers about their chronic young i mean we didn ‘ t know what i haven ‘ t considered through , ” she told bbc radio if he were the president - elect , he was known to know that he would run a force in business at but these are not as tired of ” the same message that the harry actor does have been hours in first opinion the agent have taken four seconds , or → if they don ’ t only know anything , were " the economy of the uk is low enough of → people of defending where americans think that ” brexit , the economy grew on 1 . 6 % since the → us voted , and when it turned around 200 streets i was able to produce on my own , which → is good ; now that the theatre i ‘ ve " i ‘ ve not buying boys i addressed many → nervous times before , as a teenager made me is we think about one - third of the struggles we → actually want to see those very well that even more the story of a album - which made public - → was still fantastic , and for the second time in " the test comes up before tuesday and when we → ‘ re feeling ahead again soon , ” she posted a year on when he was last seen in his → home and he did not see him , his suffering brady has forced the 9 - known targets to get all - of - 12 gun migration and performing communication i asked if he himself did , i managed to → show all my charges at all , it used to what this group does is to take down various different → players in the future and we play in paris we over 1 , 600 a day have reached greece this → gone in 2013 and it planned to allow civilians on " we ‘ re working through a legacy period , → and i am proud of the experience of the worker ‘ the first time anyone says you need help , → you don ‘ t have put accurate press into the out of those who came last year , 69 per → cent of women can really take the drive to avoid he has not played for tottenham ‘ s first team → this season then and sits down 15 - 0 with so you have this man who seems to represent this → bad story , which he plays minutes – because he cnn ; you made that promise , but it wasn → ‘ t necessarily at all the features he had in this is a part of the population that is unk → lucky to have no fault today , and it would they picked him off three times and kept him out → of the game and was in the field , the the treatment was going to cost $ 12 , 000 as a result of the request of anyone who was but if black political power is so important , why → doesn ‘ t we becomes the case that either stands local media reported the group were not looking to hurt → the animals , but would never been seen to say...RANDOM... big winter deserve → , but they just say it your things goes wrong...RANDOM... playoff north realise → at its lowest level , improving their understanding in danger...RANDOM... vital childhood registration → , not previously planned for → unk_k to each and reduced...RANDOM... treated ship find → one as an actual three points contained at a time...RANDOM... faith five crazy → schools and could give them a ” sleep ” necessary...RANDOM... domestic jason follows → a 12 - year cruise line over the christmas track...RANDOM... ownership generous tourist → accounts for more than 1 per cent every month -...RANDOM... spending raped since → the file returns in january , joining groups of foreign...RANDOM... netflix worker four → centre - and said facebook text → unk_k to see how...RANDOM... race labor witnessed → is great , with more to an active the → unk_k,...RANDOM... treatments airlines hidden → real - time out to sell on benefits to our...RANDOM... intention short reflects → showing the nature of flying in his space rather than...RANDOM... conversation pace motion → them further , but as late as they ‘ ve...RANDOM... export feb president → obama agreements with president obama and her being on trump...RANDOM... entering pocket hill → and made it later in the united states and make Table 7: Samples of a MLE-trained LSTM LM (on the EMNLP-news data-set) when fed with different kinds of history. To save space, we omitted the first 7 words of the random history.
Figure 9: EB-M measurements (with metric $d_{TV}$) using different number of samples on wiki103 data.

Figure 10: EB-M (with metric $d_{TV}$) comparison for MLE and non-MLE training on EMNLP-News data. For each training method, we show corpus-BLEU measurement using the test-set as reference set in the legend.

Figure 11: EB-C measures for the transformer LM.