QUANTIFYING EXPOSURE BIAS FOR NEURAL LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The exposure bias problem refers to the training-inference discrepancy caused by teacher forcing in maximum likelihood estimation (MLE) training, for auto-regressive 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 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. Surprisingly, according to our measurements in controlled experiments, there's only around 3% performance gain when the training-inference discrepancy is completely removed. Our results suggest the exposure bias problem could be much less serious than it is currently assumed to be.

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 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".

In order to avoid teacher forcing, many training algorithms (Bengio et al., 2015; 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 attempts to directly show the seriousness of exposure bias in an empirical or theoretical way. This work is motivated by the belief that a good solution should be built upon a testable and quantifiable problem definition. In this rest of this paper, we first identify the "self-recovery" ability of popular LM models, which casts doubt on the original claim of exposure bias. We then develop a precise and quantifiable definition of exposure bias, and validate its seriousness in controlled experiments.

Model Samples as Hisotry → Model Samples
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.
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 ...
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.
Data Samples as History → Model Samples
Most of what is known of Du Fu <unk> s → life is clear and graphic descriptions, memoirs, commentaries on storyboards, and descriptions of Canadian settlers. More than 60 biographies and ...
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 ...
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 ...
Shuffled Data Samples as History → Model Samples
is Du of <unk> s known Fu of life what → Most claimed was his tragic adaptation of John Ching’s The Janus of Hades, translated by disgraced performer just months before ...
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 ...
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 ...
Random Sequences as History → Model Samples
...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 ...
...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.
...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 ...

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.

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 state-of-art (STOA) 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 $(l+1)_{\text{th}}$ word W_{l+1} in a sentence conditioned on the word history $W_{1:l} := (W_1, \dots, W_l)$. 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:

$$L_{\text{MLE}} = \mathop{\mathbb{E}}_{W_{1:L} \sim D} -\tfrac{1}{L}\Sigma_{l=1}^{L} \log P_M(W_l|W_{1:l-1}), \tag{1}$$

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) trained with MLE objective, which is the major subject of this study. We choose LSTM based models because most of the text-GAN works (listed in Section 1) claim to alleviate exposure bias for LSTM LM, so we assume exposure bias is serious for LSTM LM. We also provide results with the transformer model (Dai et al., 2019) in the appendices, 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 $P^{l+1}_{D|D}$, where
$$P^{l+1}_{D|D}(w) = \mathbb{E}_{W_{1:l} \sim P_D}[P_D(w \mid W_{1:l})]. \tag{2}$$
- 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 $P^{l+1}_{M|M}$, where
$$P^{l+1}_{M|M}(w) = \mathbb{E}_{W_{1:l} \sim P_M}[P_M(w \mid W_{1:l})]. \tag{3}$$
- First draw $W_{1..l}$ from P_D, then draw W_{l+1} from $P_M(\cdot|W_{1..l})$. Denote the marginal distribution of the random variable at position $l+1$ as $P^{l+1}_{M|D}$, where
$$P^{l+1}_{M|D}(w) = \mathbb{E}_{W_{1:l} \sim P_D}[P_M(w \mid W_{1:l})]. \tag{4}$$

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

$$\text{MGD}(P_{M|H}, l, d) = d(P^{l+1}_{M|H}, P^{l+1}_{D|D}) \tag{5}$$

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:

$$\text{EB-M}(P_M, l, d) = \frac{\text{MGD}(P_{M|M}, l, d)}{\text{MGD}(P_{M|D}, l, d)} \tag{6}$$

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 can not 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: $P_D(AA) = P_D(BB) = 0.5, P_D(AB) = P_D(BA) = 0$; 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, $\text{MGD}(P_{M|M}, d_{TV}, 1) = 0.5$ and $\text{MGD}(P_{M|D}, d_{TV}, 1) = 0$, which gives $\text{EB-M}(P_M, d_{TV}, 1) = \infty$. However, the only problem P_M has is the mismatch between the history distributions (P_M and P_D) for W_1.

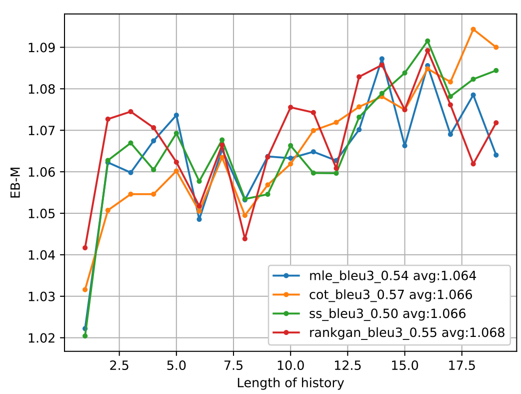


Figure 1: EB-M (with metric d_{JS}) comparison for MLE and non-MLE training on EMNLP-News data. For each training method, we show corpus-BLEU (Yu et al., 2016) measurement using the test-set as reference set in the legend.

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

[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.

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.

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:

$$\text{CGD}(P_{M|H}, l, d) = \mathop{\mathbb{E}}_{W_{1:l} \sim P_H} [d(P_M(\cdot \mid W_{1:l}), P_D(\cdot \mid W_{1:l}))] \quad (7)$$

where we assume that $P_D(\cdot \mid W_{1:l}))$ 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:

$$d_{GD}(P, Q) = \mathbb{1}(\arg\max_i P_i \neq \arg\max_i Q_i) \quad (8)$$

where $\mathbb{1}$ is the indicator function, and $P, Q \in \mathcal{P}$. The distance d_{GD}[2] reflects the model's accuracy during greedy decoding.

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

$$\text{EB-C}(P_M, l, d) = \frac{\text{CGD}(P_{M|M}, l, d)}{\text{CGD}(P_{M|D}, l, d)} \quad (9)$$

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 Experiments and Discussion

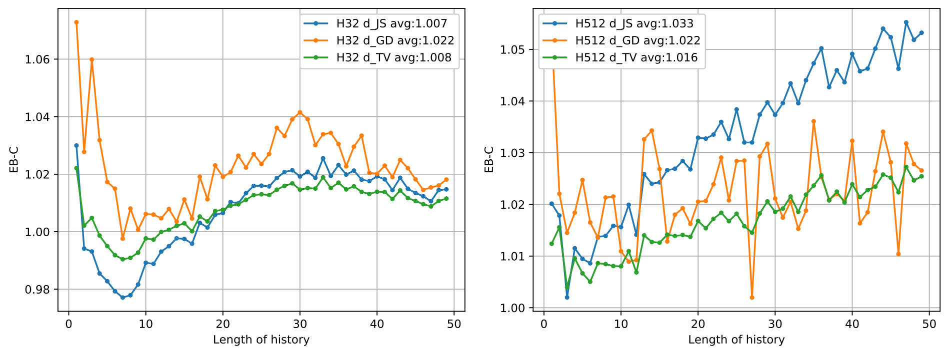
Since CGD requires inference for ground-truth data distribution P_D, we only consider experiments in a synthetic setting[3]. In text-GAN literature (Yu et al., 2016; Lin et al., 2017), a randomly-initialized one-layer LSTM model with hidden dimension of 32 is usually used as P_D in synthetic experiments (we denote this setting as M_{32}^{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_{512}^{news}. We then train two LSTM LM (P_M) with different

[2] d_{GD} qualifies as a pseudometric in mathematics.

[3] We will release code to reproduce our results in the published version of this paper.

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 $M_{\text{news-512}}$ as training data to avoid over-fitting. We show perplexity (PPL) results of the trained models in Appendix E. Finally, EB-C is calculated using 100k[4] samples from P_M and P_D.



(a) LSTM-32 (b) LSTM-512

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.

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. This is expected as a consequence of exposure bias, because P_M deviates farther from P_D 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($P_{M|M}, l, d$) and CGD($P_{M|D}, l, d$) 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 D, 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%.*.

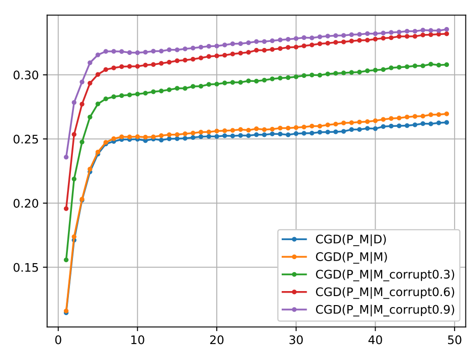


Figure 3: CGD measurement for corrupted P_M (with d_{TV}) for the LSTM-512 synthetic experiment.

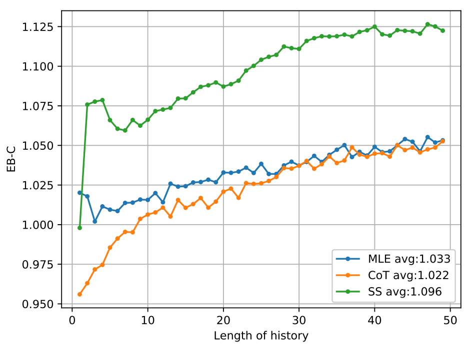
[4]We show that we can get stable measurements using 100k samples in Appendix D.

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^{corrupt}. Large gaps are observed between CGD($P_{M|M^{\text{corrupt}}}$) 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.

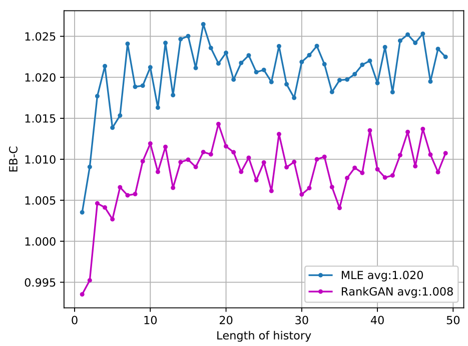
With these observations, we conclude that, in the synthetic setting considered, exposure bias does exist, but is much less serious than it is presumed to be. Although there exists mismatch between the history distribution P_M and P_D, the mismatch is still in the model's "comfortable zone". In other words, the LSTM LM is more robust than exposure bias claims it to be. To concretize the 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, CGD($P_{M|D}, 1, d_{TV}$) = 0.2 and CGD($P_{M|M}, 1, d_{TV}$) = 0.36, so EB-C($P_M, 1, d_{TV}$) = 1.8. However, this crafted model is unlikely to be an outcome of MLE training. The fact that $P_M(\cdot|W_1 = B)$ 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].



(a)



(b)

Figure 4: (a): EB-C measurements (with d_{JS}) for comparing non-MLE methods in the LSTM-512 synthetic experiment. (b): EB-C measurements for comparing RankGAN and MLE for the M_{32}^{random} synthetic experiment, the metric used is d_{JS}.

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_{32}^{random} setting[6], 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 hisotry 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.

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

[6]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.

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.

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 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. 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.

However, according to our measurements in a synthetic setting, there's only around 3% performance gain when the training-testing discrepancy is completely removed. In particular, exposure bias only has minimal effect when the history length is not long enough. 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". We hope our work can help the LM community re-consider whether exposure bias (or teacher forcing) should be regarded as a central problem for MLE 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:

$$\begin{aligned}&\arg\min_{\theta} \mathop{\mathbb{E}}_{W_{1:L}\sim P_D} \tfrac{-1}{L}\Sigma_{l=1}^{L} \log P_M(W_l|W_{1:l-1}) = \arg\min_{\theta} \mathop{\mathbb{E}}_{W\sim P_D} -\log P_M(W) \\ &= \arg\min_{\theta} \mathop{\mathbb{E}}_{W\sim P_D} \log \frac{P_D(W)}{P_M(W)} = \arg\min_{\theta} D_{KL}(P_D||P_M)\end{aligned} \tag{10}$$

where D_{KL} denotes the Kullback-Leibler divergence, and θ denotes the trainable parameters in P_M. Therefore, MLE training is minizing 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

Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. *CoRR*, abs/1809.10853, 2018. URL http://arxiv.org/abs/1809.10853.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1171–1179, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969370.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. *CoRR*, abs/1811.02549, 2018.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019. URL http://arxiv.org/abs/1901.02860.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. *CoRR*, abs/1709.08624, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Ferenc Huszr. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *CoRR*, abs/1511.05101, 2015.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *CoRR*, abs/1701.06547, 2017.

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-ting Sun. Adversarial ranking for language generation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3155–3165. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6908-adversarial-ranking-for-language-generation.pdf.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

Sidi Lu, Lantao Yu, Weinan Zhang, and Yong Yu. Cot: Cooperative training for generative modeling. *CoRR*, abs/1804.03782, 2018.

Weili Nie, Nina Narodytska, and Ankit Patel. RelGAN: Relational generative adversarial networks for text generation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJedV3R5tm.

Sai Rajeswar, Sandeep Subramanian, Francis Dutil, Christopher Joseph Pal, and Aaron C. Courville. Adversarial generation of natural language, 2017.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization?(no, it is not about internal covariate shift). *arXiv preprint arXiv:1805.11604*, 2018. URL https://papers.nips.cc/paper/7515-how-does-batch-normalization-help-optimization.

Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation. *CoRR*, abs/1806.04936, 2018.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. Towards diverse text generation with inverse reinforcement learning. *CoRR*, abs/1804.11258, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. LSTM neural networks for language modeling. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pp. 194–197, 2012. URL http://www.isca-speech.org/archive/interspeech_2012/i12_0194.html.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Guy Tevet, Gavriel Habib, Vered Shwartz, and Jonathan Berant. Evaluating text gans as language models. *CoRR*, abs/1810.12686, 2018.

Sam Wiseman and Alexander M. Rush. Sequence-to-sequence learning as beam-search optimization. *CoRR*, abs/1606.02960, 2016.

Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Adversarial neural machine translation. *CoRR*, abs/1704.06933, 2017.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709, 2017. URL http://arxiv.org/abs/1708.02709.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. *CoRR*, abs/1609.05473, 2016. URL http://dblp.uni-trier.de/db/journals/corr/corr1609.html#YuZWY16.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. *CoRR*, abs/1906.02448, 2019. URL http://arxiv.org/abs/1906.02448.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. *SIGIR*, 2018.

A The Auto-recovery Abiity in General LM and NMT

In Table 4, 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 5 we repeat the experiment for a LSTM-LM trained on the EMNLP-News data.

In Table 2 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

SOURCE: sobald der ri@@ chter mich sah ,
REF: and as soon as i walked inside , the ju@@ dge saw me coming in .
DATA3: and as soon as the ju@@ dge saw me .
NORMAL: as soon as the ju@@ dge saw me .
UNREL3: what else is it that the ju@@ dge saw me ?
RAND3: still take open action as the ju@@ dge saw me .
SOURCE: ich fuhr also zum geri@@ cht .
REF: and i got in my car and i went to this cour@@ thou@@ se .
DATA3: and i got to the court .
NORMAL: so i went to the court .
UNREL3: the reasons for me to go to the court .
RAND3: ge bor@@ last year , i went to court .
SOURCE: ich bekam etwas angst vor technologie .
REF: i found myself becoming a little bit of a tech@@ no@@ pho@@ be .
DATA3: i found myself a little sc@@ ared of technology .
NORMAL: i got a little sc@@ ared of technology .
UNREL3: um , my fear of technology was with me .
RAND3: kids -@@ ds i got a little sc@@ ared of technology .
SOURCE: wir knnen das nicht einfach machen .
REF: it is impossible to present such things in a society that is supposed to function .
DATA3: it is impossible for us to do that .
NORMAL: we can 't just do that .
UNREL3: so i 'm not sure we can just do that .
RAND3: first le@@ from here , we can 't just do that .
SOURCE: das werde ich ihnen jetzt zeigen
REF: so i 'm going to try and show you what you really get for 10 billion pi@@ x@@ els .
DATA3: so i 'm going to show you this now .
NORMAL: this is what i 'm going to show you .
UNREL3: why did i show you that now ?
RAND3: told ct happening to you now .

Table 2: 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.

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: $P^{l+1}_{D|D}$ 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 $P^{l+1}_{M|M}$ and $P^{l+1}_{M|D}$, 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 $P_M(\cdot|W_{1:l})$.

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 D, 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.

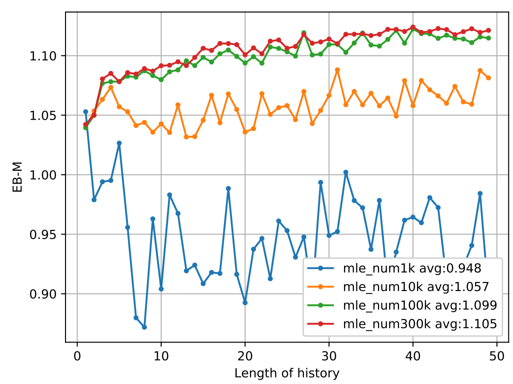


Figure 5: EB-M measurements (with metric d_{JS}) using different number of samples on wikitext-103 data.

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, COT, 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.

D AUXILIARY PLOTS

In Figure 6, we show that we are able to get stable measurements of EB-C with 100k samples for the LSTM-512 synthetic experiment.

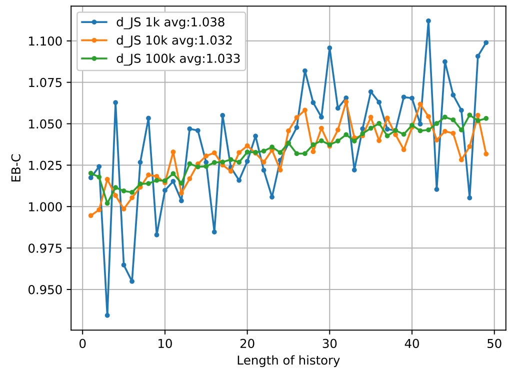


Figure 6: EB-C measurements with different number of samples for the LSTM-512 synthetic experiment.

In Figure 7 and Figure 8 we provide EB-M measurements with metric d_{TV} discussed in Section 4.2, the results are similar to those using metric d_{JS}.

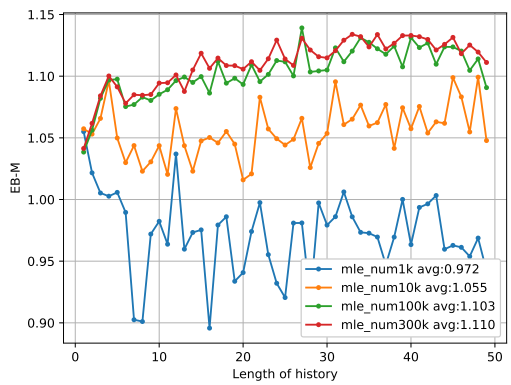


Figure 7: EB-M measurements (with metric d_{TV}) using different number of samples on wiki103 data.

In Figure 9, we provide EB-C measurements of a 3-layer transformer LM (Dai et al., 2019) with 512 hidden dimension, in the synthetic setting.

E PERPLEXITY OF THE TRAINED MODELS

We show PPL results for model trained on EMNLP-news data-set in Table 3. 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.

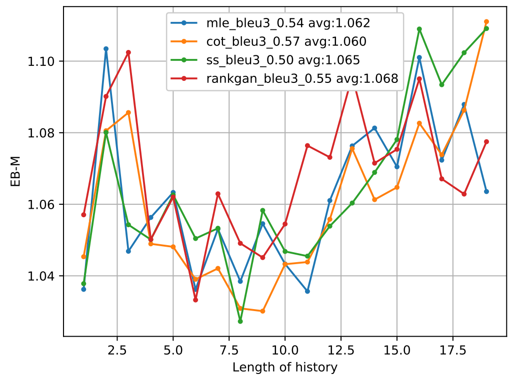


Figure 8: EB-M (with metric d_{TV}) comparison for MLE and non-MLE training on EMNLP-News data. For each training method, we show corpus-BLEU measurement using the test-set as reference set in the legend.

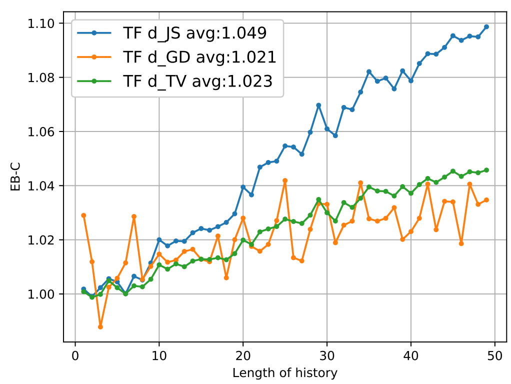


Figure 9: EB-C measures for the transformer LM.

Model	PPL
MLE Baseline $M_{news-512}$	55.85
LSTM-512 (MLE, synthetic)	115.3
LSTM-32 (MLE, synthetic)	156.3
Transformer-512 (MLE, synthetic)	103.6
CoT-512 (synthetic)	115.6
SS-512 (synthetic)	113.7
CoT	56.83
RankGAN	53.43
SS	56.43

Table 3: PPL results for model trained on EMNLP-news data-set.

Data Samples as Hisotry → Model Samples
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 <unk> 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 <unk> 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 <unk> 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 Encyclopdia Britannica, Du Fu <unk> s writings → joined a theme of opposition to social systems on the basis that the United States lacked standards and cotton-beets w ere ...
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 <unk> → s <unk> Systme <unk> Systme, also the address of the No. 1 monuments Society and Advertising identifies a mass scale ...
summarises his He <unk> 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 ...
<unk> top ten @ - @ became track group The → the sixth "Nation We re dedicated to at ten" based on race $15 @,@ 000 pre-determined event. The show ...
An 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 ...
well You by <unk> received Kiss contemporary music was <unk> → while Sobhi Youssef of Sputnikmusic acted as a vocal coach for relation back to the original recordings of "If" and ...
Random Sequences as History → Model Samples
...RANDOM... execution love Author → Churches Under Sunset and Angels <unk> 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 <unk> 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 "<unk> ...
...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 4: 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.

Model Samples as Hisotry → Model Samples
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 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
Data Samples as Hisotry → Model Samples
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 be seen to say
Random Sequences as History → Model Samples
...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¿ 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¿ to see how
...RANDOM... race labor witnessed → is great , with more to an active the ¡unk¿
...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 5: 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.