Blank Language Models

Tianxiao Shen[1] **Victor Quach**[1] **Regina Barzilay**[1] **Tommi Jaakkola**[1]

Abstract

We propose Blank Language Model (BLM), a model that generates sequences by dynamically creating and filling in blanks. Unlike previous masked language models (Devlin et al., 2018) or the Insertion Transformer (Stern et al., 2019), BLM uses blanks to control which part of the sequence to expand. This fine-grained control of generation is ideal for a variety of text editing and rewriting tasks. The model can start from a single blank or partially completed text with blanks at specified locations. It iteratively determines which word to place in a blank and whether to insert new blanks, and stops generating when no blanks are left to fill. BLM can be efficiently trained using a lower bound of the marginal data likelihood, and achieves perplexity comparable to traditional left-to-right language models on the Penn Treebank and WikiText datasets. On the task of filling missing text snippets, BLM significantly outperforms all other baselines in terms of both accuracy and fluency. Experiments on style transfer and damaged ancient text restoration demonstrate the potential of this framework for a wide range of applications.[1]

1. Introduction

Neural language models have been successfully applied to many sequence generation tasks, including machine translation (Bahdanau et al., 2014), summarization (Rush et al., 2015), and image captioning (Xu et al., 2015). Typically, sequences are modeled autoregressively from left to right, making the log-likelihood tractable and allowing efficient training and inference. While left to right models are effective, they are not well-suited for text completion or editing. In these tasks, we are given a partial draft of the text and the goal is to add new text to complete it.

Models such as Masked Language Model (Devlin et al., 2018, MLM) and Insertion Transformer (Stern et al., 2019)

---

[1]MIT CSAIL. Correspondence to: Tianxiao Shen <tianxiao@csail.mit.edu>.

[1]Our code will be released soon.

They also have _____ which _____ .
They also have _ice cream_ which _is really good_ .

Figure 1. BLM fills in blanks of arbitrary length.

are able to fill in words to complete partially written text. However, neither of them is tailored to rewriting/editing. MLM assumes that the length of the text to be inserted is known in advance. Insertion Transformer, on the other hand, does not explicitly control where insertions can take place.

In this paper, we introduce Blank Language Model (BLM). The model exploits a special "__" symbol to control where tokens can be placed. In each stage of generation, a blank can be replaced by any word, and potentially accompanied by a new blank on the left, right or both sides of the word to continue writing. As shown in Fig. 1, such models can be used to fill in missing words in incomplete sentences, generate a new sentence in between two given sentences, and so on. BLM can start with a single blank or partial text with blanks in specified locations. The model iterates through generation steps, replacing blanks with words and possibly adjoining blanks, until no blanks remain.

Our BLM is based on a Transformer encoder that maps the input text containing blanks into a sequence of vector representations. The representations at blank locations are further processed to select a blank, word to fill in it, and whether to generate adjoining blanks. Since there are multiple trajectories through the actions in the BLM that all result in the same final text, we train the model by maximizing the marginal likelihood. To make training more efficient, and to introduce an inductive bias towards order independence, we maximize instead a lower bound on the marginal likelihood. At test time, BLM can in principle fill in any amount of text in any of the given blank positions.

We test BLM on language modeling, and obtain perplexity comparable to left-to-right language models on Penn Treebank and WikiText datasets. We further evaluate our model on three text rewriting tasks: text infilling (Zhu et al., 2019), ancient text restoration (Assael et al., 2019) and style transfer (Shen et al., 2017). BLM achieves superior performance on all three tasks, demonstrating its flexibility to generate text in diverse conditions. Notably, on ancient text restoration, we reduce the previous state-of-the-art error rate from 44.9% to 41.6% when half of the characters are missing.

	Canvas		Action			
Step			Location b	Word w	(Left l,	Right r)
0.	#1		#1	is	Y	Y
1.	#1 is #2		#1	customer	N	Y
2.	customer #1 is #2		#2	awesome	N	N
3.	customer #1 is awesome		#1	service	N	N
4.	customer service is awesome			-End-		

Figure 2. An example trajectory that generates the sentence "customer service is awesome". Each action is a tuple (b, w, l, r), indicating the blank location b selected for expansion, the word w to fill in, whether to create a left blank l, and whether to create a right blank r.

2. Related Work

Alternatives to conventional left-to-right generation have previously been explored from multiple approaches. Part of these efforts was focused on finding an optimal generation order, including syntax-based approaches and methods for learning adaptive generation order (Emami & Jelinek, 2005; Zhang et al., 2015; Dyer et al., 2016; Ford et al., 2018; Zhou et al., 2019; Welleck et al., 2019; Gu et al., 2019a). These approaches are tailored to generation from scratch in a specific order. Our model instead is attuned for text rewriting, where the missing parts can be located anywhere in the input text, and the algorithm must flexibly complete them.

Another stream of work focuses on generating sequences in a non-autoregressive fashion for fast decoding in machine translation (Gu et al., 2017; Lee et al., 2018; Stern et al., 2019; Gu et al., 2019b). The closest approach is the Insertion Transformer (Stern et al., 2019), which also supports a dynamic canvas growing with word insertions. However, none of these models provide explicit control over which part of the sequence to expand.

Additional insertion control is provided by the masked language model where each mask corresponds to a single word (Fedus et al., 2018). MLMs are commonly used in representation learning (Devlin et al., 2018). To utilize them in rewriting tasks would require one to specify the insertion length in advance and heuristically determine a generation order among masks (Ghazvininejad et al., 2019; Wu et al., 2019). In contrast, a blank in our model can correspond to any number of words, thereby avoiding the problem of predicting length. BLMs provide a natural formulation for generative modeling that can dynamically accommodate insertions of various length.

Finally, several works combine left-to-right language models with control codes or customized inference algorithms for more flexible generation (Keskar et al., 2019; Sun et al., 2017; Liu et al., 2019). Our model allows for straightforward decoding strategies and enables direct edits to the sentence to control generation.

3. Blank Language Models

A blank language model (BLM) generates sequences by creating and filling in blanks. Generation starts with a single blank and ends when there is no blank. In each step, the model selects a blank "__", predicts a word w, and fills the blank with "w", "__ w", "w __", or "__ w __". In this way, a blank can be expanded to any number of words.

We define a *canvas* as a sequence of words interspersed with special "__" tokens. The subsequent action is conditioned on this intermediate stage of generation. Different from the Insertion Transformer that can insert words anywhere in between existing tokens (Stern et al., 2019), the BLM will only place words on the specified blanks.

Suppose the current canvas is $c = (c_1, \cdots, c_n)$ with blanks located at indices $b_1, \cdots, b_k$ (i.e. $c_{b_l} = $ "__", for $l = 1, \ldots, k$). BLM maps this canvas to a distribution over actions specifying how the canvas is to be revised:

$$p(b, w, l, r | c; \theta) = \text{BLM}(c) \qquad (1)$$

where $b \in \{b_1, \cdots, b_k\}$ is a blank location; w is a word in the vocabulary V; $l, r \in \{0, 1\}$ denote whether or not to create a blank to the left and right of w; and θ are the model parameters. The *action*, defined as the tuple (b, w, l, r) uniquely specifies the next state of canvas (see Fig. 2).

We can view the actions in BLM alternatively as production rules in a grammar. Each blank represents a nonterminal symbol (or the start symbol), and the terminal symbols come from the vocabulary V. The production rules are restricted to be of the form "__" → "__ $?w$ __?" for $w \in V$, where "?" indicates that the preceding symbol is optional. In contrast to context free grammars, the probability distribution over production rules is conditioned on the entire canvas generated so far.

Model Architecture To implement the model, we first encode $(c_1, \cdots, c_n)$ into a sequence of representations $(z_1, \cdots, z_n)$, and then take corresponding representations $z = (z_{b_1}, \cdots, z_{b_k})$ where the blanks are located. Let d represent the dimension of z. We factorize the joint distribution into three parts (see Fig. 3 for an overview):

1) Choose a blank

Linear & Softmax

z_0 ($b_i=0$) z_1 ($b_i=1$)

Transformer

They   also   have   ___   which   .

positional embedding 방식?

2) Predict a word

Linear & Softmax → really

3) Create new blanks

MLP →

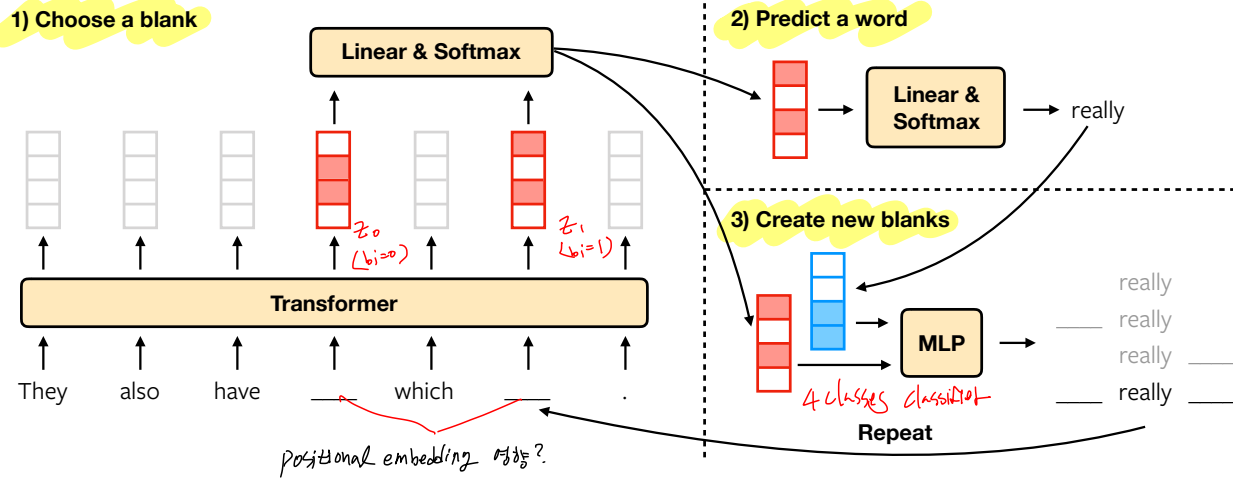4 classes classifier

really
really
really
really

Repeat

Figure 3. Architecture of the Blank Language Model. In the first stage, an index is chosen among all current blank positions. For that location, a word is selected in the second stage. In the final stage, the blank representation is concatenated with the chosen word's embedding and fed into a multilayer perceptron (MLP) to determine the creation of the following blanks.

Model Architecture   To implement the model, we first encode $(c_1, \cdots, c_n)$ into a sequence of representations $(z_1, \cdots, z_n)$; and then take corresponding representations $z = (z_{b_1}, \cdots, z_{b_k})$ where the blanks are located. Let d represent the dimension of z. We factorize the joint distribution into three parts (see Fig. 3 for an overview):

1. Choose a blank:

$$p(b_i|c; \theta) = \text{Softmax}(zu) \quad (2)$$

where $u \in \mathbb{R}^d$ is a parameter vector to project z's into one-dimensional logits.

z: Representation set of blank

2. Predict a word for the selected blank:

$$p(w|c, b_i; \theta) = \text{Softmax}(W z_{b_i}) \quad (3)$$

where $W \in \mathbb{R}^{|V| \times d}$ is a parameter matrix to project z_{b_i} into the vocabulary.

3. Decide whether or not to create blanks to the left and right of the predicted word:

$$p(l, r|c, b_i, w; \theta) = \text{MLP}(z_{b_i}, v_w) \quad (4)$$

where v_w is the word vector of w, and MLP is a multilayer perceptron network with 4 output classes: (Left/Right) × (Yes/No).

Likelihood   Now let us consider the probability $p(x; \theta)$ of generating a sentence/paragraph x under the BLM. We call the generating process from an initial blank to complete text a *trajectory*. The same final text x may be realized by multiple trajectories. However, if we specify the order in which the words in x are generated, the trajectory is also uniquely determined. This follows from the fact that BLM never results in a canvas with two (or more) consecutive blanks. Concretely, consider the example trajectory of a 4-word sentence in Fig. 2. Given the order $(3, 1, 4, 2)$, at step 0 when we generate x_3, we must create both left and right blanks for future generations of x_1 and x_2, x_4. In step 1 of generating x_1, we create a right blank but no left blank

because there are no more words on x_1's left. Subsequent steps can be deduced by analogy. The correspondence between trajectories and generation orders allows us to write the marginal likelihood as:

$$p(x; \theta) = \sum_{\sigma \in S_n} p(x, \sigma; \theta) = \sum_{\sigma \in S_n} \prod_{t=0}^{n-1} p(a_t^{x,\sigma}|c_t^{x,\sigma}; \theta) \quad (5)$$

where S_n is the set of all n-permutations, and $a_t^{x,\sigma}, c_t^{x,\sigma}$ denote the action and canvas at step t respectively (cf. Fig. 2).

Training   Previous work explored different strategies to design loss for training generalized sequence models (Devlin et al., 2018; Yang et al., 2019; Stern et al., 2019; Chan et al., 2019). Here we propose training objectives derived from log likelihood.

Directly computing the marginal likelihood over $n!$ orders is intractable. We use Jensen's inequality to lower bound the log likelihood: https://en.m.wikipedia.org/wiki/Jensen%27s_inequality

$$\log p(x; \theta) = \log \sum_{\sigma \in S_n} \prod_{t=0}^{n-1} p(a_t^{x,\sigma}|c_t^{x,\sigma}; \theta)$$

$$\geq \log(n!) + \frac{1}{n!} \sum_{\sigma \in S_n} \log \prod_{t=0}^{n-1} p(a_t^{x,\sigma}|c_t^{x,\sigma}; \theta)$$

$$= \log(n!) + \frac{1}{n!} \sum_{\sigma \in S_n} \sum_{t=0}^{n-1} \log p(a_t^{x,\sigma}|c_t^{x,\sigma}; \theta)$$

$\forall x_i \in (a,b), \; p_i > 0, \; \sum_{i=1}^{n} p_i = 1$

$f\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i f(x_i)$

$$(6)$$

where equality holds when the posterior $p(\sigma|x; \theta)$ is uniform. By maximizing this lower bound we also minimize the gap between the true posterior and the uniform approximation. As a result, we impose an inductive bias towards

x: A B C D

S_n: (1, 2, 3, 4)

(2, 1, 3, 4)

(3, 1, 2, 4)

...

	Canvas		Action			
Step			Location b	Word w	(Left l,	Right r)
0.	#1		#1	is	Y	Y
1.	#1	is #2	#1	customer	N	Y
2.	customer #1 is #2		#2	awesome	N	N
3.	customer #1 is awesome		#1	service	N	N
4.	customer service is awesome			-End-		

Text infilling	Input:	They also have ____ which ____ .
	Target:	They also have <u>ice cream</u> which <u>is really good</u> .
Ancient text restoration	Input:	τε εγγονον εισαι<u>???????</u>σοφιαι
	Target:	τε εγγονον εισαι<u>ου του</u>σοφιαι
Style transfer	Positive:	The employees behind the deli counter were **super nice** and **efficient** !
	Negative:	The employees behind the deli counter were <u>rude</u> and <u>unprofessional</u> !

Figure 4. Examples of inputs and outputs for the three rewriting tasks. We contrast text infilling, where blanks __ can cover an arbitrary number of words, with ancient text restoration, where the number of characters to recover is indicated by the number of '?' symbols in the input.

learning to realize x equally well, independent of the order. This is desirable to ensure that the model is able to complete any partial input text regardless of the position of the blanks.

From Equation (6), we can derive our first (naive) training algorithm. First, sample a permutation σ from S_n and a step t from 0 to $n-1$, then compute the estimated loss $[-\log(n!) - n \cdot \log p(a_t^{x,\sigma}|c_t^{x,\sigma};\theta)]$. However, this procedure has a large variance and can only compute the loss of a single action in one pass (in contrast to left-to-right language models that compute n word losses per pass).

To train more efficiently, we note that the canvas $c_t^{x,\sigma}$ depends only on the first t elements of σ. Hence we can combine loss calculations of trajectories that are the same in the first t steps but different at the $t+1$ step. Switching the summation order of σ and t, we have:

$$\frac{1}{n!}\sum_{t=0}^{n-1}\sum_{\sigma \in S_n} \log p(a_t^{x,\sigma}|c_t^{x,\sigma};\theta)$$

$$= \sum_{t=0}^{n-1}\sum_{\sigma_{1:t}}\sum_{\sigma_{t+1}} \frac{(n-t-1)!}{n!} \cdot \log p(a_t^{x,\sigma}|c_t^{x,\sigma};\theta) \quad (7)$$

This leads to our efficient training algorithm: first sample t and $\sigma_{1:t}$, then construct the canvas $c_t^{x,\sigma}$, and compute loss $\left[-\log(n!) - \frac{n}{n-t}\sum_{\sigma_{t+1}} \log p(a_t^{x,\sigma}|c_t^{x,\sigma};\theta)\right]$. In this way, we can compute in expectation $n/2$ action losses per pass.

4. Experiments

We start by measuring the performance of BLM on language modeling benchmarks and comparing it with traditional left-to-right language models as a sanity check. We then demonstrate the BLM's ability to rewrite specified portions of text in a document by evaluating it on three text editing tasks: text infilling (Zhu et al., 2019), ancient text restoration (Assael et al., 2019) and style transfer (Shen et al., 2017). Figure 4 displays example inputs and outputs for these tasks.

Experimental Details In all experiments, the sequence representations in BLM are obtained using the encoder module of a `transformer_base` architecture (Vaswani et al.,

	PTB	WT2	WT103
LSTM (Grave et al., 2016)	82.3	99.3	48.7
TCN (Bai et al., 2018)	88.7	-	45.2
Transformer (Dai et al., 2019)	-	-	30.1
Adaptive (Baevski & Auli, 2018)	-	-	18.7
Transformer-XL (Dai et al., 2019)	54.5	-	18.3
Insertion (Stern et al., 2019)	77.3	91.4	55.8
BLM (ours)	69.2	81.2	58.2

Table 1. Perplexity on the Penn Treebank and WikiText datasets.

2017) (6 layers, 8 heads, $d_{model} = 512$, $d_{ff} = 2048$, $d_k = d_v = 64$). The MLP network used for blank prediction has one hidden layer of size 1024. Weight decay, learning rate and dropout are tuned based on the perplexity achieved on the validation set. For tasks that require decoding, we use beam size in $\{1, 5, 10, 20\}$ and choose the best value as observed on the validation set. We note that beam search in BLM does not search for the sentence with the maximum marginal likelihood $p(x;\theta)$, but instead for a sentence *and* a trajectory that have the maximum joint likelihood $p(x,\sigma;\theta)$.

4.1. Language Modeling

To compute the perplexity of the BLM and the Insertion Transformer, we use the Monte-Carlo method to estimate the likelihood in Eq. (5) with $m = 1000$ samples.

Datasets We test on three benchmark datasets: Penn Treebank (PTB) (Mikolov et al., 2010), WikiText-2 (WT2) and WikiText-103 (WT103) (Merity et al., 2016).

Results Table 1 summarizes the perplexity of our model in comparison with previous work. The top results are achieved by the Transformer-XL (Dai et al., 2019) and the adaptive embedding method (Baevski & Auli, 2018). However, these systems have additional advantages in terms of utilization of supplementary techniques and their model size. On WikiText-103, the models in Dai et al. (2019) and Baevski & Auli (2018) use 250M parameters, whereas
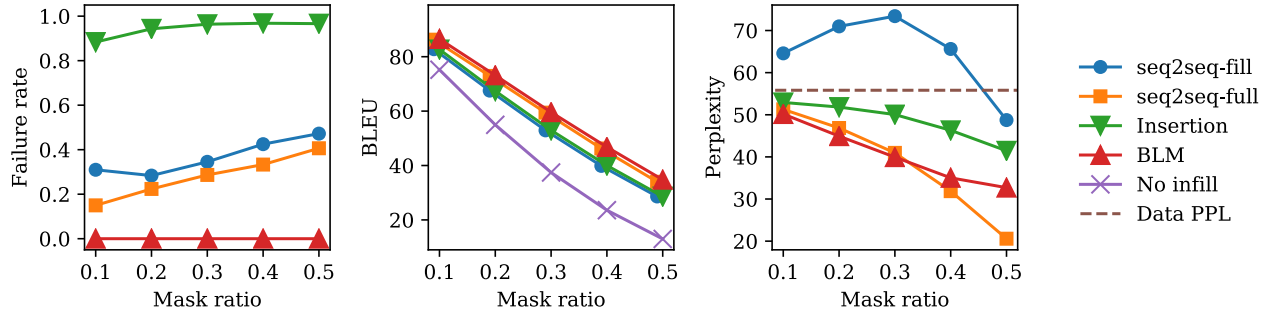
Figure 5. Failure rate, BLEU score and perplexity of generated documents for the text infilling task. The "No infill" line reports the BLEU score of the blanked document. The "Data PPL" dotted line serves as reference for the perplexity of the original documents

our model uses 42M parameters. These advancements can also be combined with our model. When evaluated against comparable baselines, our model rivals the Insertion Transformer as well as left-to-right language models with LSTM and Temporal Convolutional Network (TCN) architecture. The finding is particularly noteworthy, since the language modeling task is more challenging for free-order models like ours.

4.2. Text Infilling

The task of text infilling is motivated by many practical applications where the goal is to augment partially completed documents with missing information (Zhu et al., 2019). Following the protocol of Zhu et al. (2019), we automatically compile test data by deleting portions of documents, and ask systems to fill them in. The first row in Fig. 4 showcases an example input-output pair. The infilling task evaluates model's ability to complete blanks in a document while maintaining semantic consistency with the imposed context.

Dataset We experiment on the Yahoo Answers dataset (Yang et al., 2017), which has 100k training documents and 10k documents for validation and testing respectively. Each document has 78 words on average. For a document x, we randomly mask a given ratio r of its tokens. Contiguous masked tokens are collapsed into a single blank token "__", resulting in a canvas c with k such blanks. The systems are required to complete the blanks in c.

Baselines We compare our approach against the following three baselines:

- The *seq2seq-full* baseline is a Transformer model trained to output the full document x from input c. Note that it may have invalid outputs that do not match the input format, such as missing existing tokens in c or generating tokens in incorrect locations.

- The *seq2seq-fill* baseline is a Transformer model that only generates tokens to be placed in the blanks, with a special '|' token to indicate separation. For the example in Fig. 4, its target output will be "ice cream |is really good". Unlike *seq2seq-full*, *seq2seq-fill* does not have the problem of losing existing tokens in c. However, it may still fail to generate the correct number of '|' tokens that matches the input.

- The *Insertion Transformer* does not explicitly support controlling the position of insertion. We force it to generate words only in the designated blanks by normalizing the predictions over valid locations. Note that the model still may not fill all of the required blanks.

Metrics Following prior work (Zhu et al., 2019; Liu et al., 2019), we measure the accuracy of generation by computing its BLEU score against the original document x, and the fluency of generation as its perplexity evaluated by a pre-trained (left-to-right) language model. In addition, we report the failure rate of baselines, defined as the percentage of invalid generations, i.e. generations that do not respect the constraints of the task.

Results In Figure 5, we plot the failure rate, BLEU score, and perplexity of models at different mask ratios. Our BLM is the only method that is able to consistently generate valid outputs. Seq2seq baselines have a failure rate ranging from 15% to 56% as the mask ratio increases. Insertion Transformer has the highest failure rate: in more than 88% of cases, it does not fill all the blanks. This indicates that the Insertion Transformer is not suitable for generation with location constraints.

According to the BLEU score, BLM and *seq2seq-full* have the highest infilling accuracy, on average 5.8 points higher than that of the Insertion Transformer and *seq2seq-fill*. For reference, we also plot the BLEU score of the input canvas

	Mask-ratio 0.1	Mask-ratio 0.5
Blanked	when time flies , ____ does it go ? ____ the center of the ____ to be recycled ____ made into new time .	when time ____ , where ____ ? ____ the ____ of ____ universe to ____ recycled ____ made into ____ .
BLM	when time flies , *where* does it go ? *for* the center of the *earth* to be recycled *and* made into new time .	when time *was created* , where *did it come from* ? *it was* the *first part* of *the* universe to *be* recycled *and* made into *space* .
Insertion	when time flies , *where* does it go ? _ the center of the *earth has* to be recycled *and* made into new time .	when time *was created* , where _ ? the *name* of *the* universe to *be* recycled *and* made into *space* .
seq2seq-full	when time flies , *where* does it go ? *at* the center of the *earth* to be recycled *and* made into new time .	when time *heals* , where *does it go* ? *it 's* the *end* of *the* universe to *be* recycled *and* made into *space* .
seq2seq-fill	when time flies , *how* does it go ? *at* the center of the *earth* to be recycled *and* made into new time . how |at |earth |and	when time *is time* , where *is time* ? *time is* the *time* of *time* universe to *the* recycled *be* made into *and* . the universe is time |is time |time is |time |time |the |be |and |the universe
Original	when time flies , **where** does it go ? **to** the center of the **universe** to be recycled **and** made into new time .	when time **flies** , where **does it go** ? **to** the **center** of **the** universe to **be** recycled **and** made into **new time** .

Figure 6. Example generations for the text infilling task, with mask ratios 0.1 and 0.5. Completions are in italic. Invalid completions are in red. For the *seq2seq-fill* baseline, we represent the outputs of the model along with the merged document. In this example, the insertion transformer produces invalid completions by failing to generate tokens in the "? __ the" blank. At mask ratio 0.5, the *seq2seq-fill* baseline also generates an invalid document by producing too many '|' tokens, i.e. filling to many blanks.

with respect to the original document. When the mask ratio is 0.5, the input BLEU score is 13.0, and BLM brings it up to 34.8 after infilling. In terms of fluency, with the exception of seq2seq-fill, the outputs of all other methods have perplexity lower than the original data perplexity. This is because with greedy decoding or beam search, the models tend to generate the most typical output with the highest likelihood.

The inspection of typical generations validates the superiority of BLM. In Fig. 6, we present an illustrative output for each model at different mask ratios. In the low mask ratio setting, models only need to use a single word to fill in blanks and produce a grammatically correct completion. Most models successfully accomplish this task. With the higher mask ratio of $r = 0.5$ where half of the words are deleted and the main ideas of the document are concealed, the infilling task is much more challenging and requires models to creatively generate sentences that fit the imposed canvas. Although the original meaning of the sentence is not recovered, BLM is the only model able to produce a coherent document with consistency between the question and the answer.

Overall, BLM displays the best performance both quantitatively and qualitatively. For seq2seq approaches, generating the full document is superior to generating only the infilled content. Probably because that in the former case the decoder can better model the full text, whereas in the latter case the decoder must model segmented text and meanwhile count for blanks.

4.3. Ancient Text Restoration

Ancient text restoration is a form of text infilling where there exist fragments in ancient documents that are illegible due to time-related damages and need to be recovered (Assael et al., 2019). The second row in Figure 4 illustrates an example of input and output for the task. Restoration is performed at the character-level, and the number of characters to recover is assumed to be known, denoted by a '?' symbol in the input. In reality, when epigraphists restore a deteriorated document, the length of the lost fragment is unknown and needs to be guessed as a first step. While previous work relies on these expert conjectures (Assael et al., 2019), we note that our formulation is able to bypass this limitation and can flexibly generate completions without this additional knowledge. For purposes of comparison, however, we evaluate our method on the length-aware setting.

Length-aware Blank Language Model (L-BLM) We present a variant of the BLM that is well-suited to the specific features of this task. The vocabulary V is an alphabet of characters from the ancient Greek language. We extend the vocabulary V with special " __[t]__ " tokens that denote the length of the fragment to recover. Specifically, as a preprocessing step, consecutive '?' characters are collapsed into a single " __[t]__ " token, where t is the number of '?' symbols. For each such blank token, L-BLM is trained to predict a character and the lengths of the new blanks to its left and right. In all experiments, we use special blank tokens for lengths up to 1000 and follow our usual canvas creation procedure.

Mask ratio	Single-slot 1%	Multi-slot 25%	40%	50%
HUMAN	57.3%	-	-	-
PYTHIA	32.5%	-	-	-
PYTHIA-WORD	**29.1%**	**36.9%**	42.3%	44.9%
L-BLM (ours)	33.7%	37.1%	**37.9%**	**41.6%**

Table 2. Character error rate for the ancient text restoration task in both single-slot and multi-slot settings.

Dataset The PHI-ML dataset (Assael et al., 2019) is made of fragments of ancient Greek inscriptions containing more than 3 million words and 18 millions characters. We evaluate models in two settings: *single-slot* and *multi-slot*. The test set is generated following Assael et al. 2019's procedure: a context of length $L = 1000$ is sampled from an inscription, then a slot of length $C \in [1, 10]$ is sampled from that context. The characters from that slot are replaced with the '?' prediction symbol and constitute the target. For the single-slot experiment, we use the testing script from prior work (Assael et al., 2019) and sample 12,800 testing samples, for a total of 63,234 characters to predict, with mask ratio of 1.2%. For the multi-slot setting, we progressively increase the number of slots, yielding larger mask ratios. In total, we generate a total of 1000 samples for each mask ratio of 25%, 40% and 50% with respectively 150,235, 400,827 and 406,231 characters to restore.

Baselines Previous work has proposed PYTHIA (Assael et al., 2019), a sequence-to-sequence based approach specialized in ancient text restoration. A variant of PYTHIA, PYTHIA-WORD, uses both character and word representation as input. During training, the model learns to recover masked characters using examples where a single slot has been sampled, with a slot length limited to 10. For the multi-slot setting, PYTHIA is applied iteratively as described in Assael et al. 2019. Beam search of size 20 is applied to each independent prediction.

Metrics We measure the character error rate (CER) of all models in both settings.

Results Table 2 summarizes the experimental results. L-BLM achieves similar character error rate as PYTHIA in the single-slot setting, significantly outperforming human experts. When PYTHIA is augmented with word representations, the model is able to further decrease the error rate compared to character-only methods.

In reality, restoring damaged inscriptions requires the reconstruction of multiple lost fragments. As a larger proportion of the text is removed, PYTHIA-WORD's performance is degraded. In contrast, L-BLM is robust to this setting change and significantly outperforms prior work. We posit that L-BLM's advantage lies in its ability to efficiently maximize

the joint likelihood of the completions over all slots. In contrast, PYTHIA-WORD's is only aware of one slot at a time. Moreover, L-BLM can handle slots of arbitrary long length while PYTHIA-WORD is limited to slots of up to 10 characters, which is a limiting factor for real-world usage.

4.4. Sentiment Transfer

The goal of sentiment transfer is to modify the sentiment of a sentence while maintaining its topic (Shen et al., 2017). An example is described on the third row of Figure 4. Inspired by the way humans perform rewriting, we follow a recent line of work in style transfer (Li et al., 2018; Xu et al., 2018; Wu et al., 2019) that adopts a two-step approach:

1. Remove words and expressions of high polarity from the source sentence;

2. Complete the partial sentence with words and expressions of the target sentiment.

Step 1 has been performed in previous work by masking tokens either based on their frequency-ratio (Li et al., 2018; Wu et al., 2019) or their attention scores (Xu et al., 2018; Wu et al., 2019). Step 2 is performed by various sequence models conditioning on the masked sentence and the target sentiment.

We evaluate the contribution of our model in Step 2 as a substitute for infilling models used in prior pipelines (Li et al., 2018; Wu et al., 2019). To this end, we train two instances of BLM on the dataset, one for each sentiment. At test time, the corresponding BLM is used to produce completions of the target sentiment.

Dataset We run experiments on the benchmark Yelp review dataset (Shen et al., 2017), using the standard split of 450K non-parallel training sentences, 4K validation sentences and 1K testing sentences. Each sentence is labeled as either positive or negative.

Baselines We compare the performance of our model against two infilling methods. The DELETE-AND-RETRIEVE method (Li et al., 2018) is a seq2seq-based approach where hidden representations of the masked sentence is concatenated with a learned attribute embedding before decoding. Additionally, a retrieval module is used to collect relevant expressions of the target sentiment to guide generation. The MASK-AND-INFILL model (Wu et al., 2019) is based on a pretrained BERT_{base} model and then fine-tuned by conditioning on the sentiment of the sentence to reconstruct.

Metrics We use evaluation methods introduced by prior work (Shen et al., 2017; Li et al., 2018; Wu et al., 2019; Yang et al., 2018). To assess the accuracy of the generated sentences with respect to the target sentiment, we use a

	ACC (%)	BLEU
w/frequency-ratio		
Canvas (mask only)	33.9	21.9
DELETE-AND-RETRIEVE	**88.3**	13.4
MASK-AND-INFILL (MLM)	58.0	18.7
BLM (ours)	64.5	**23.3**
w/attention-based		
Canvas (mask only)	42.6	19.9
MASK-AND-INFILL (MLM)	40.0	21.8
BLM (ours)	**79.6**	**21.9**

Table 3. Accuracy and BLEU scores for the Yelp sentiment transfer task. Accuracy measures the percentage of sentences labeled as the target sentiment by the classifier. BLEU is evaluated against human reference generations. For reference, we also report accuracy and BLEU scores of the canvas (i.e. the original masked sentence).

pretrained CNN classifier that achieves 97.7% accuracy on the validation set. We also measure the BLEU score between the transferred sentences and human references (Li et al., 2018).

Results Results in Table 3 demonstrate the ability of different models to perform text infilling for style transfer. The DELETE-AND-RETRIEVE method with the frequency-ratio based masking strategy achieves high sentiment accuracy, but can only do so at the expense of content fidelity. By constraining BLM to fill in blanks in between content words, we ensure that the predictions will yield high content preservation, improving both BLEU score and sentiment accuracy over the original masked sentence.

The MLM formulation in MASK-AND-INFILL is problematic on this task for two reasons. By design, MLM is forced to generate the same number of tokens as there were originally in the source sentence, making it more difficult to produce coherent sentences that are consistent with the target sentiment. Furthermore, MLM is trained to predict the masked tokens independently rather than jointly, which further hurts performance. Our formulation of BLM does not suffer any of these weaknesses. With both masking strategies, our model outperforms the MASK-AND-INFILL baseline on all metrics, proving its superiority as the better-suited formulation for this setup[2].

In Fig 7, we present examples generated by the blank language model. BLM is able to dynamically adapt to the imposed canvas and can fill in blanks with expressions of varied lengths, such as "**very helpful**" → "rude" or "**nowhere to be** found" → "the best i found". We note that failure cases arise when negative polarity items are left unmasked;

---

[2]Wu et al. 2019 experiment with using a classifier loss to constrain training and report that it can dramatically improve the MLM's generations accuracy. We suspect similar gains in performance if applied to BLM.

Source	the food 's ok , the service is **among** the **worst** i have encountered .
BLM	the food 's ok , the service is *probably* the *best* i have encountered .
Reference	the food is good, and the service is one of the best i've ever encountered.
Source	the beans were in the burro in the rice was **nowhere to be** found .
BLM	the beans were in the burro in the rice was *the best i* found .
Reference	the beans were in the burro and the rice was plentiful
Source	everyone that i spoke with was **very helpful** and **kind** .
BLM	everyone that i spoke with was *rude* and *unprofessional* .
Reference	everyone that i spoke with wasn't helpful or kind.
Source	everything is **fresh** and so **delicious** !
BLM	everything is *horrible* and so *expensive* !
Reference	everything was so stale
Source	there is **definitely not** enough **room** in that part of the venue .
BLM	there is *always* enough *parking* in that part of the venue .
Reference	there is so much room in that part of the venue
Source	it is n't **terrible** , but it is **n't** very good either .
BLM	it is n't *fancy* , but it is *still* very good either .
Reference	it is n't perfect , but it is very good .
Source	executive chefs would **walk** by **not** even saying good morning .
BLM	executive chefs would *come* by *without* even saying good morning .
Reference	the exceculive chef was nice and said good morning to us very often

Figure 7. Example generations for the style transfer task using attention-based masking mechanism. Masked words are in bold.

BLM is then unable to produce satisfactory outputs from the canvas.

5. Conclusion

In this paper, we proposed the blank language model for flexible text generation. BLMs can generate sequences in different orders by dynamically creating and filling in blanks. We demonstrate the effectiveness of our method on various text rewriting tasks, including text infilling, ancient text restoration and style transfer. Future work may explore sequence modeling tasks beyond text rewriting that also benefit from flexible generation order. An example is music modeling: harmonic constraints naturally impose a canvas that composers fill in with the melody.

References

Assael, Y., Sommerschield, T., and Prag, J. Restoring ancient text using deep learning: a case study on Greek epigraphy. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6368–6375, 2019.

Baevski, A. and Auli, M. Adaptive input representations for neural language modeling. *arXiv preprint arXiv:1809.10853*, 2018.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bai, S., Kolter, J. Z., and Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Chan, W., Kitaev, N., Guu, K., Stern, M., and Uszkoreit, J. Kermit: Generative insertion-based modeling for sequences. *arXiv preprint arXiv:1906.01604*, 2019.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. Recurrent neural network grammars. *arXiv preprint arXiv:1602.07776*, 2016.

Emami, A. and Jelinek, F. A neural syntactic language model. *Machine learning*, 60(1-3):195–227, 2005.

Fedus, W., Goodfellow, I., and Dai, A. M. Maskgan: better text generation via filling in the _. *arXiv preprint arXiv:1801.07736*, 2018.

Ford, N., Duckworth, D., Norouzi, M., and Dahl, G. E. The importance of generation order in language modeling. *arXiv preprint arXiv:1808.07910*, 2018.

Ghazvininejad, M., Levy, O., Liu, Y., and Zettlemoyer, L. Constant-time machine translation with conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

Gu, J., Liu, Q., and Cho, K. Insertion-based decoding with automatically inferred generation order. *Transactions of the Association for Computational Linguistics*, 7:661–676, 2019a.

Gu, J., Wang, C., and Zhao, J. Levenshtein transformer. In *Advances in Neural Information Processing Systems*, pp. 11179–11189, 2019b.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Lee, J., Mansimov, E., and Cho, K. Deterministic non-autoregressive neural sequence modeling by iterative refinement. *arXiv preprint arXiv:1802.06901*, 2018.

Li, J., Jia, R., He, H., and Liang, P. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *NAACL*, 2018.

Liu, D., Fu, J., Liu, P., and Lv, J. TIGS: An inference algorithm for text infilling with gradient search. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4146–4156, 2019.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., and Khudanpur, S. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pp. 6830–6841, 2017.

Stern, M., Chan, W., Kiros, J., and Uszkoreit, J. Insertion transformer: Flexible sequence generation via insertion operations. *arXiv preprint arXiv:1902.03249*, 2019.

Sun, Q., Lee, S., and Batra, D. Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6961–6969, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*. 2017.

Welleck, S., Brantley, K., Daumé III, H., and Cho, K. Non-monotonic sequential text generation. *arXiv preprint arXiv:1902.02192*, 2019.

Wu, X., Zhang, T., Zang, L., Han, J., and Hu, S. Mask and infill: Applying masked language model for sentiment transfer. In *IJCAI*, 2019.

Xu, J., Sun, X., Zeng, Q., Zhang, X., Ren, X., Wang, H., and Li, W. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 979–988, 2018.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Yang, Z., Hu, Z., Salakhutdinov, R., and Berg-Kirkpatrick, T. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3881–3890. JMLR. org, 2017.

Yang, Z., Hu, Z., Dyer, C., Xing, E. P., and Berg-Kirkpatrick, T. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pp. 7287–7298, 2018.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5754–5764, 2019.

Zhang, X., Lu, L., and Lapata, M. Top-down tree long short-term memory networks. *arXiv preprint arXiv:1511.00060*, 2015.

Zhou, L., Zhang, J., and Zong, C. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, 7:91–105, 2019.

Zhu, W., Hu, Z., and Xing, E. Text infilling. *arXiv preprint arXiv:1901.00158*, 2019.