Sentence-Permuted Paragraph Generation

Wenhao Yu†, Chenguang Zhu‡, Tong Zhao†, Zhichun Guo†, Meng Jiang†
†University of Notre Dame ‡Microsoft Research
†{wyu1, tzhao2, zguo5, mjiang2}@nd.edu
‡ chezhu@microsoft.com

Abstract

Generating paragraphs of diverse contents is important in many applications. Existing generation models produce similar contents from homogenized contexts due to the fixed left-to-right sentence order. Our idea is permuting the sentence orders to improve the content diversity of multi-sentence paragraph. We propose a novel framework *PermGen* whose objective is to maximize the expected log-likelihood of output paragraph distributions with respect to all possible sentence orders. *PermGen* uses hierarchical positional embedding and designs new procedures for training, decoding, and candidate ranking in the sentence-permuted generation. Experiments on three paragraph generation benchmarks demonstrate *PermGen* generates more diverse outputs with a higher quality than existing models.

1 Introduction

Paragraph generation is an important yet challenging task. It requires a model to generate informative and coherent long text that consists of multiple sentences from free-format sources such as a topic statement or some keywords (Guo et al., 2018). Typical paragraph generation tasks include story generation (Fan et al., 2018), news generation (Leppänen et al., 2017), scientific paper generation (Koncel-Kedziorski et al., 2019), etc. Recent advances in natural language generation models such as Transformer (Vaswani et al., 2017) and BART (Lewis et al., 2020) have demonstrated attractive performance of generating text paragraphs.

An important desired property of model-generated paragraphs is diversity – given the same source, an intelligent model is expected to create a variety of paragraphs in terms of content, semantic style, and word variability (Li et al., 2016; Ippolito et al., 2019). For example, a story generation model should narrate a plot with different storylines (Clark et al., 2018); a scientific paper generation model should suggest diverse contents to spark new ideas (Wang et al., 2019). In order to create diversity, controllable methods (Zhao et al., 2017; Cho et al., 2019; Yu et al., 2020) used additional inputs (e.g., aspects, styles). Sampling decoding algorithms (Radford et al., 2019; Holtzman et al., 2020) searched next tokens widely from a vocabulary. However, existing models struggled to produce multi-sentence paragraphs of diverse contents, because they relied on the homogeneity of contexts (e.g., similar story beginnings) caused by the conventional autoregressive framework with fixed left-to-right sentence order (i.e., S1→S2→S3).

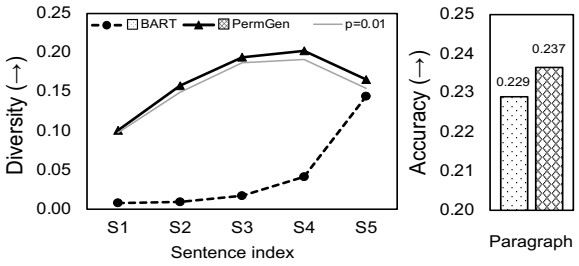

Figure 1: Left: Diversity of each generated story sentence at different positions (1 to 5) in ROCStories' test set, measured by averaged 1-Self-BLEU (Zhu et al., 2018). Our *PermGen* produces contents of higher diversity at all positions, while BART (dashed line) produces diverse outputs only at the end of story. With p-value<0.01, *PermGen* has higher diversity than the the grey line. Right: *PermGen* outperforms BART in the accuracy of generated stories measured by BLEU-4.

As an example, Figure 1 evaluates the diversity of each generated sentence at different positions of the story in ROCStories (Mostafazadeh et al., 2016) by different models. As shown, BART (dashed line) tends to generate stories of very similar beginning and middle parts and only produce diverse text near the end of a story. This phenomenon stems from the fact that the left-to-right generation leads

§ Our code and output files are available at https://github.com/wyu97/permgen.

to homogeneity of context to the left, reducing the diversity of the generated paragraph.

Our idea is permuting the sentence orders in paragraph generation, while sticking with the left-to-right scheme to generate tokens in each sentence. It has two advantages. First, it provides an output sentence with a variety of contexts (and possibilities) from different orders. For example, creating the story ending first can probably produce a completely different story from generating the beginning first. Second, it retains the benefit of autoregressive model that originates from the word-by-word nature of human language production. So the coherence within sentences can be maintained, avoiding the harm of incomplete semantics from token-level permutation (Shen et al., 2020).

In this work, we propose a sentence-permuted paragraph generation framework called *PermGen*. Instead of using the fixed forward order, *PermGen* maximizes the expected log-likelihood of the distribution in output paragraph *w.r.t.* all possible sentence orders. The optimization is based on π-SGD (Murphy et al., 2019) which has guaranteed convergence property. Furthermore, *PermGen* employs a novel hierarchical position encoding scheme to represent the positions of tokens in permuted sentences. *PermGen* can be initialized with any Transformer-based models and any decoding algorithms such as beam search and nucleus sampling (Holtzman et al., 2020).

We conduct experiments on three paragraph generation tasks: story generation, news generation, and paper abstract generation. Results show that *PermGen* can significantly improve the diversity of generated texts and achieve higher accuracy. Particularly, as shown in Figure 1, *PermGen* model can improve diversity for sentences at all positions while also improving the accuracy. Besides, we observe consistent improvements on both accuracy and diversity when *PermGen* is coupled with various pre-trained models and decoding algorithms.

2 Related Work

Paragraph Generation. The source can be either structured or unstructured such as database records (Puduppully et al., 2019), knowledge graphs (Zhao et al., 2020), images (Ippolito et al., 2019), and keywords (Yao et al., 2019). The expected outputs typically are stories (Guan et al., 2019; Yao et al., 2019), essays (Yang et al., 2019), news articles (Leppänen et al., 2017), or scientific papers (Hua and Wang, 2019; Koncel-Kedziorski et al., 2019). This task poses unique challenges as it aims at generating coherent and diverse long-form texts. Our framework can use various forms of input such as a story title, keywords, and keyphrases, which can be generalized to broad domains.

Diverse Text Generation. Generating diverse sequences is of crucial importance in many text generation applications that exhibit semantically *one-to-many* relationships between source and the target sequences, such as machine translation (Shen et al., 2019; Lachaux et al., 2020), summarization (Cho et al., 2019), question generation (Wang et al., 2020), and paraphrase generation (Qian et al., 2019). Methods of improving diversity in text generation that have been widely explored from different perspectives in recent years. Sampling-based decoding is one of the effective solutions to improve diversity (Fan et al., 2018; Holtzman et al., 2020), e.g., nucleus sampling (Holtzman et al., 2020) samples next tokens from the dynamic nucleus of tokens containing the vast majority of the probability mass, instead of aiming to decode text by maximizing the likelihood. Another line of work focuses on introducing random noise (Gupta et al., 2018) or changing latent variable (Lachaux et al., 2020) to produce uncertainty, e.g., Gupta et al. (2018) employ a variational auto-encoder framework to generate diverse paraphrases according to the input noise. In addition, Shen et al. (2019) adopt a deep mixture of experts (MoE) to diversify machine translation, where a minimum-loss predictor is assigned to each source input; Shi et al. (2018) employ inverse reinforcement learning for unconditional diverse text generation.

Dynamic Order Generation. These methods have two categories. First, non-autoregressive generation is an emerging topic and commonly used in machine translation (Gu et al., 2018; Ren et al., 2020). They generate all the tokens of a sequence in parallel, resulting in faster generation speed. However, they perform poorly for long sentences due to limited target-side conditional information (Guo et al., 2019). Second, insertion-based generation is a partially autoregressive model that maximizes the entropy over all valid insertions of tokens (Stern et al., 2019). POINTER (Zhang et al., 2020) inherits the advantages from the insertion operation to generate text in a progressive coarse-to-fine manner. Blank language model (BLM) (Shen et al., 2020) provides a formulation for generative modeling that

accommodates insertions of various length.

Different from the above methods, our *PermGen* permutes the sentence orders for generating a paragraph, and it follows the left-to-right manner when producing each sentence.

3 Problem Definition

Given input X that can be a topic statement, some keywords, or a paper's title, the goal is to produce a paragraph Y consisting of multiple sentences as a story, a news article, or a paper's abstract. Suppose Y has T sentences, denoted by $Y = [Y_1, \cdots, Y_T]$, where Y_t is the t-th sentence. T can be easily obtained from training data to create sentence indices. During testing, models are expected to predict the sentence indices under maximum T (i.e., 10).

3.1 Sentence-Level Transformer

Transformer (Vaswani et al., 2017) follows the encoder-decoder architecture (Sutskever et al., 2014) and uses stacked multi-head self-attention and fully connected layers for both the encoder and decoder. For simplicity, we represent the Transformer framework at the sentence level by using a recurrent notation that generates a probability distribution for sentence prediction by attending to both input X and previous decoded sentences $Y_{<t}$.

$$p(Y_t) = \text{Transformer}(X, Y_{<t}). \quad (1)$$

where Y_t and $Y_{<t}$ are the t-th sentence and sentences *before* t-th sentence under the left-to-right manner in target output. Transformer eschews recurrence and instead relies on the self-attention mechanism to draw global dependencies between the input and output. During the decoding phase, Transformer can predict each token based on both the input and previously predicted tokens via attention masks to improve efficiency. The objective of Transformer is to maximize the likelihood under the forward autoregressive factorization:

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(Y_t|Y_{<t}, X;\theta). \quad (2)$$

4 Proposed Method: *PermGen*

In a left-to-right generation scheme such as the canonical Seq2Seq design, each generated token is conditioned on left-side tokens only (Sutskever et al., 2014). It ignores contextual dependencies from the right side. It also leads to limited diversity of generated text (as shown in Figure 1). To solve this problem, our *PermGen*, a novel sentence-permuted paragraph generation model, produces sentences not confined to the left-to-right order. Instead, *PermGen* attempts different sentence orders and selects the best-ranked output candidate.

As shown in Figure 2, *PermGen* uses the Transformer encoder but changes the sentence orders during the decoding phase. It should be noted that *PermGen* follows the left-to-right manner when generating tokens in each sentence. Thus, we represent the Transformer decoder as:

$$Y_{\pi_t} = \text{Transformer}(X, Y_{\pi_{<t}}, \pi), \quad (3)$$

where Y_{π_t} and $Y_{\pi_{<t}}$ are the t-th sentence and the sentences *before* the t-th sentence under the permuted order π in the target output. Taking the first permuted order in Figure 2 as an example, we have $\pi = [2, 1, 3]$, $\pi_1 = 2$, $\pi_3 = 3$, $\pi_{<3} = [2, 1]$.

We note that as *PermGen* is based on the encoder-decoder Transformer architecture, which can be initialized either randomly or from a pre-trained Transformer model with the same structure. For example, in the experiments, we evaluate *PermGen* which is i) trained from scratch, and ii) initialized with BART (Lewis et al., 2020). Next, we will introduce three modules of *PermGen*: (1) hierarchical positional embedding, (2) sentence-permuted learning, and (3) sentence-based decoding.

4.1 Hierarchical Positional Embedding

In Transformer, positional embeddings are added to every token's embedding. Traditionally, the positional embedding encodes the absolute position from 1 to the sequence length to model how a token at one position attends to tokens at other positions (Vaswani et al., 2017; Lewis et al., 2020).

We propose the hierarchical positional embedding that consists of a global position and a local position. Given a token, the global position is the position (index) of the sentence that contains this token; the local position is the position of the token in the sentence (see the two lines of position numbers in Figure 2). Given a paragraph Y, its embedding matrix is given below, where rows are its tokens and columns are embedding dimensions:

$$\mathbf{Y} = \mathbf{Y}_{\text{token}} + \mathbf{Y}_{\text{global\_position}} + \mathbf{Y}_{\text{local\_position}}, \quad (4)$$

where $\mathbf{Y}_{\text{token}}$ is the token embedding, $\mathbf{Y}_{\text{global\_position}}$ and $\mathbf{Y}_{\text{local\_position}}$ are the global positional embeddings and local positional embeddings.

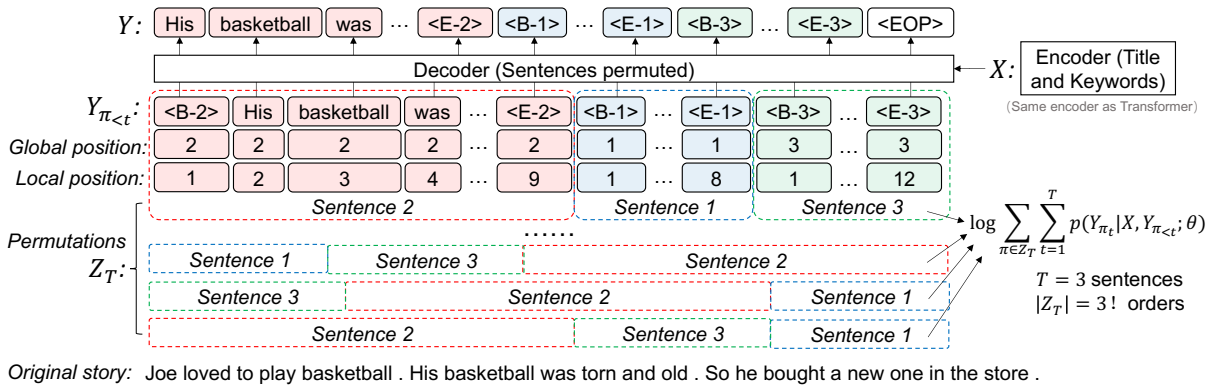

Figure 2: The architecture of *PermGen*. The example story has 3 sentences, leading to $3! = 6$ permuted sentence orders. *PermGen* minimizes the overall generation loss w.r.t. all possible sentence orders.

Compared to the absolute positional embedding, the hierarchical positional embedding has two advantages. First, the embedding of two-level positions is more informative about the paragraph structure than that of the absolute position. Second, when we permute the sentence orders in paragraph generation, the absolute positions of tokens might not be available. For example, if the second sentence is generated earlier than the first sentence, the absolute positions of its tokens cannot be determined because the length of the first sentence is unknown. In comparison, hierarchical position does not have this issue.

In addition, for the t-th sentence in Y, we add two special tokens (i.e., <B-t> and <E-t>) to indicate the beginning and end of the sentence. Thus, the decoder can determine the sentence index based on the predicted special tokens. We also append a special token <EOP> to the paragraph to indicate the end of the generation process.

4.2 Sentence-permuted Learning

This module learns by varying sentence orders in paragraph generation and acts as the key component in *PermGen*. For example, given a sentence order $\pi = [2, 4, 1, 5, 3]$, *PermGen* first generates the second sentence from the leftmost token to the rightmost, then generates the fourth sentence, and so on. The model stops when the third sentence is finished. Formally, we denote Z_T as the set of all possible sentence orders, i.e., the permutations of sentence indices of length T. It follows that $|Z_T| = T!$. Given input X and target output paragraph Y of T sentences, *PermGen* maximizes the following likelihood:

$$p(Y|X;\theta) = \sum_{\pi \in Z_T} p(Y|X,\pi;\theta) = \sum_{\pi \in Z_T} \prod_{t=1}^{T} p(Y_{\pi_t}|X, Y_{\pi_{<t}};\theta) \quad (5)$$

However, computing the negative log-likelihood in Eq. (5) is prohibitive because the back-propagation computational graph branches out for every permutation in the sum. Therefore, we apply the Jensen's inequality to lower-bound the log-likelihood:

$$\log p(Y|X;\theta) = \log \sum_{\pi \in Z_T} \prod_{t=1}^{T} p(Y_{\pi_t}|X, Y_{\pi_{<t}};\theta)$$

$$\geq \log(|Z_T|) + \frac{1}{|Z_T|} \sum_{\pi \in Z_T} \sum_{t=1}^{T} \log p(Y_{\pi_t}|X, Y_{\pi_{<t}};\theta)$$

By maximizing the lower bound, we do not favor any particular sentence order, but encourage the model to generate Y equally well in all orders.

Note that maximizing this lower bound is equivalent to minimizing the following expectation:

$$\mathcal{J}(\theta) = \mathbb{E}_{\pi'}\Big[-\sum_{t=1}^{T} \log p(Y_{\pi'_t}|X, Y_{\pi'_{<t}};\theta)\Big]. \quad (6)$$

Although computing this expectation is still intractable, we apply the π-SGD (Murphy et al., 2019) stochastic optimization, which randomly samples a permutation for gradient computation each time.

Definition 1 *(π-SGD): Let $\mathcal{B} = \{(X^{(1)}, Y^{(1)}), \ldots, (X^{(B)}, Y^{(B)})\}$ be a mini-batch i.i.d. sampled uniformly from the training data D. At step t, consider the stochastic gradient descent update*

$$\theta_t = \theta_{t-1} - \eta_t G_t, \quad (7)$$

where $G_t = -\frac{1}{B}\sum_{i=1}^{B} \nabla_\theta \sum_{t=1}^{T} \log p(Y^{(i)}|X^{(i)}, \pi';\theta)$ is the gradient, and random permutations $\{\pi'_i\}_{i=1}^{B}$ are sampled independently: $\pi'_i \sim$

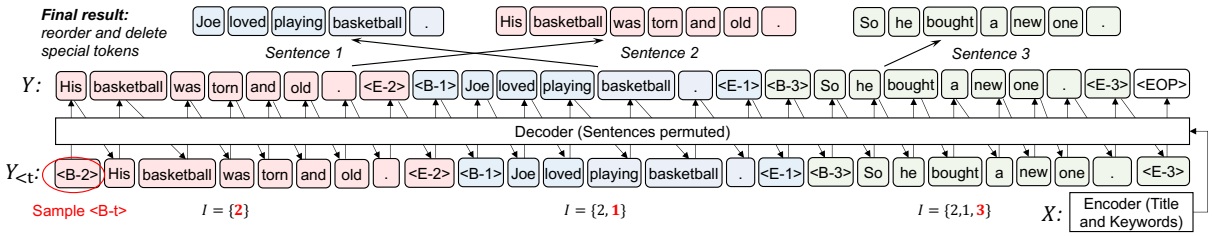

Figure 3: The decoding process during inference as described in Section 4.3. Note that the first special token (e.g., <B-2>) is sampled from $\{\text{<B-}t\text{>}\}_{t=1}^{T}$. For simplicity, positional embedding is omitted in the figure.

$\text{Uniform}(Z_T^{(i)})$. *Besides, the learning rate is* $\eta_t \in (0,1)$ *s.t.* $\lim_{t\to\infty}\eta_t = 0$, *and* $\sum_{t=1}^{\infty}\eta_t^2 < \infty$.

We note that π-SGD is a Robbins-Monro stochastic approximation of gradient descent (Robbins and Monro, 1951). When it's applied to permutation sampling, the optimization almost surely converges to the optimal θ, as implied by the following proposition.

Proposition 1 *(π-SGD Convergence): The optimization of π-SGD converges to the optimal θ for $\mathcal{J}(\theta)$ in Eq. (6) with probability one.*

Proof: We refer to Prop.2.2 in Murphy et al. (2019).

4.3 Sentence-based Decoding

In decoding, *PermGen* adopts the following steps:

- *Step 1:* Initialize a set of indices of sentences that have been generated: $I = \{\}$;
- *Step 2:* If $I = \{\}$, sample a token from $\{\text{<B-}t\text{>} \mid t \in \{1, \ldots, T\}\}$[1]; otherwise, predict a token from $\{\text{<B-}t\text{>} \mid t \in \{1, \ldots, T\} \backslash I\} \cup \{\text{<EOP>}\}$. If the token is <EOP>, end; otherwise, append <B-t> to the generated text;
- *Step 3:* Generate tokens from $\mathcal{V} \cup \{\text{<E-}t\text{>}\}$ for the t-th sentence in an autoregressive way, where $\mathcal{V}$ is the set of normal text tokens. Stop when <E-t> is generated;
- *Step 4:* $I \leftarrow I \cup \{t\}$, then go back to Step 2.

As stated in *step 2*, when <EOP> is generated, the whole generation ends. Then, the sentences in the generated paragraph can be reordered according to sentence indices I and special tokens. Note that in *step 3*, since *PermGen* adopts autoregressive generation, it can employ any decoding strategy such as beam search or sampling algorithm (e.g. truncated sampling (Fan et al., 2018), nucleus sampling (Holtzman et al., 2020)). For example, truncated sampling samples the next word from the top k probable choices, instead of aiming to decode text by maximizing the likelihood.

Rank with log-probability. We compute the log-likelihood of each candidate as the same as in beam search (Vijayakumar et al., 2016) and sampling methods (Holtzman et al., 2020):

$$S_{\text{prob}}(Y) = \frac{1}{L}\sum_{l=1}^{L}\log p(y_l|y_1, \cdots, y_{l-1}) \quad (8)$$

where L is the total number of tokens in Y and y_l is the l-th token in generated paragraph Y.

Complexity reduction. Since the number of possible sentence orders grows as $n!$ for a n-sentence paragraph, exact inference is an extremely time consuming process. To reduce the complexity during inference, we employ an approximate inference by taking advantage of the special token prediction mentioned in *step 2*. The special token prediction happens when a end-of-sentence (i.e., <E-t>) is generated. Instead of traversing each remaining possible sentence index, the model only chooses the most likely sentence index through special token predictions. It should be noted that we reuse the classifier in decoder by simply masking tokens *not in* $\{\text{<B-}t\text{>}\}_{t=1}^{T}$, without training any new classifiers. Therefore, the decoding time is roughly linear in the number of candidates to be generated. See empirical analysis in Section 5.5.4.

5 Experiments

We conduct experiments on three text generation tasks: story generation, news generation, and paper abstract generation. For all tasks, we compare *PermGen* with multiple baseline models on diversity and accuracy of their generated texts. We also perform human evaluation on story generation.

5.1 Tasks and Benchmarks

Task 1: Story generation In this task, models learn to generate story paragraphs from the title and multiple keywords. We use ROCStories

[1]When trying to generate multiple candidates, we use the sampling without replacement strategy. For example, if we need to generate 3 candidates each with 5 sentences, their beginning tokens can be B-1, B-3 and B-4, respectively.

Table 1: Statistics of three datasets. "in/out" stands for input/output and "sents" stands for sentences.

Dataset	ROCStories	AGENDA	DailyMail
# Train	98,162	38,720	49,102
# Dev.	9,817	1,000	2,000
# Test	9,803	1,000	2,000
Title in input	√	√	×
Avg.in.words	9.65	16.09	7.91
Avg.out.words	50.16	76.12	95.62
Avg.out.sents	4.92	3.08	3.88

dataset (Mostafazadeh et al., 2016) and follow the same data preparation as in Yao et al. (2019). ROCStories has 98,162 / 9,817 / 9,803 paragraphs for training / development / test sets. The stories in the corpus capture causal and temporal commonsense relations between daily events.

Task 2: Paper abstract generation In this task, models need to generate paper abstracts from paper title and a list of keywords. We use the AGENDA dataset (Koncel-Kedziorski et al., 2019) that consists of 40,720 paper titles and abstracts in the Semantic Scholar Corpus taken from the proceedings of 12 AI conferences. Each abstract is paired with several keywords. We follow the settings in Koncel-Kedziorski et al. (2019) to directly generate paper abstracts from the keywords. We follow the same data partition, which has 38,720 / 1,000 / 1,000 for training / development / test sets, respectively.

Task 3: News generation In this task, models are trained to generate news articles from a list of keyphrases. We use DailyMail dataset (See et al., 2017), a corpus of online news articles. We randomly sample 53,102 news articles and extract keyphrases from each sentence using RAKE (Rose et al., 2010). It contains 49,102 / 2,000 / 2,000 news articles for training / development / test sets.

5.2 Baseline Methods

We compare with three pre-trained Transformer-based models: BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and BERTGen (Rothe et al., 2020). These models have demonstrated state-of-the-art performance in various tasks. We also compare with GPT-2 (Radford et al., 2019) and two recent non-autoregressive generation models: BLM (Shen et al., 2020) and POINTER (Zhang et al., 2020).

BLM (Shen et al., 2020) Blank Language Model (BLM) generates sequences by dynamically creating and filling in blanks. The blanks control which part of the sequence to fill out, making it ideal for word-to-sequence expansion tasks.

POINTER (Zhang et al., 2020) POINTER operates by progressively inserting new tokens between existing tokens in a *parallel* manner. This procedure is recursively applied until a sequence is completed. This coarse-to-fine hierarchy makes the generation process intuitive and interpretable.

For each task, we evaluate *PermGen* with three diversity promoting methods for decoding including Beam search, Truncated sampling (Fan et al., 2018), and Nucleus sampling (Holtzman et al., 2020). For each method, we select the top-3 candidate paragraphs for comparison.

Truncated Sampling (Fan et al., 2018) It randomly samples words from top-k candidates of the distribution at the decoding step.

Nucleus Sampling (Holtzman et al., 2020) It avoids text degeneration by truncating the unreliable tail of the probability distribution, sampling from the dynamic nucleus of tokens containing the vast majority of the probability mass.

Implementation details is in Appendix 5.3.

5.3 Implementation Details

We use pre-trained parameters from BART-base (Lewis et al., 2020) to initialized our model, which takes a maximum 512 input token sequence and consists of a 6-layer transformer encoders and another 6-layer transformer decoders (Vaswani et al., 2017) with 12 attention heads and 768 word dimensions. For model fine tuning, we use Adam with learning rate of 3e-5, L2 weight decay of 0.01, learning rate warm up over the first 10,000 steps, and linear decay of learning rate. Our models are trained with a 4-card 32GB memory Tesla V100 GPU, and implemented on PyTorch with the Huggingface's Transformer (Wolf et al., 2020).

5.4 Evaluation Metrics

We use metrics introduced in previous work (Ott et al., 2018; Vijayakumar et al., 2018; Zhu et al., 2018) to evaluate accuracy and diversity.

5.4.1 Accuracy metrics

Top-1 metric (⇑). This measures the Top-1 accuracy among the generated hypotheses. The accuracy is measured using corpus-level metrics, including BLEU (Papineni et al., 2002),

Table 2: Diversity ("Dist-2": Distinct-2(⇑), "Self-B-4": Self-BLEU-4(⇓)) and accuracy ("B-4": BLEU-4(⇑)) for *PermGen* and baseline methods. Diversity evaluation is calculated by top-k generated candidates from *beam search*. More evaluation results (e.g., METEOR, CIDEr, Entropy) are in Table 7 in Appendix.

Methods	Pre-Train	ROCStories			AGENDA			DailyMail		
		Diversity		Accuracy	Diversity		Accuracy	Diversity		Accuracy
		Dist-2(⇑)	Self-B-4(⇓)	B-4(⇑)	Dist-2(⇑)	Self-B-4(⇓)	B-4(⇑)	Dist-2(⇑)	Self-B-4(⇓)	B-4(⇑)
POINTER	√	0.0743	0.9405	0.0492	0.1898	0.9267	0.0379	0.1228	0.9619	0.0243
BLM	√	0.0560	0.9573	0.1477	0.1465	0.9396	0.1679	0.0831	0.9889	0.1164
GPT-2	√	0.0915	0.9194	0.0726	0.1665	0.9331	0.1247	0.1577	0.9287	0.1072
BERTGen	√	0.0672	0.9456	0.1576	0.1463	0.9356	0.1462	0.1167	0.9774	0.1728
T5	√	0.0684	0.9403	0.1895	0.1323	0.9421	0.1688	0.1086	0.9779	0.1529
Transformer	×	0.0806	0.9341	0.1809	0.1489	0.9265	0.1540	0.1109	0.9678	0.1496
BART	√	0.0839	0.9330	0.2445	0.1697	0.9278	0.1922	0.1306	0.9720	0.1935
PermGen	×	0.0992	0.8548	0.1848	0.2203	**0.5679**	0.1678	0.1934	0.7757	0.1592
	√	**0.1059**	**0.7993**	**0.2482**	**0.2492**	0.5940	**0.2059**	**0.2065**	**0.6627**	**0.1991**

METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015).

Oracle metric (⇑). This measures the highest accuracy comparing the best hypothesis among the top-K with the target (Ott et al., 2018; Vijayakumar et al., 2018). Concretely, we generate hypotheses $\{\hat{Y}^{(1)}, \cdots \hat{Y}^{(K)}\}$ from each source X and keep the hypothesis $\hat{Y}^{\text{best}}$ that achieves the best sentence-level metric with the target Y. Then we calculate a corpus-level metric with the greedily-selected hypotheses $\{Y^{(i),\text{best}}\}_{i=1}^{N}$ and references $\{Y^{(i)}\}_{i=1}^{N}$.

5.4.2 Diversity metrics

Corpus diversity (⇑). Distinct-k (Li et al., 2016) measures the total number of unique k-grams normalized by the total number of generated k-gram tokens to avoid favoring long sentences. Entropy-k (Zhang et al., 2018) reflects how evenly the empirical k-gram distribution is for a given sentence when word frequency is taken into account (i.e. low weights for high-frequency words).

Pairwise diversity (⇓). Referred as "self-" (e.g., self-BLEU) (Zhu et al., 2018), it measures the within-distribution similarity. This metric computes the average of sentence-level metrics between all pairwise combinations of hypotheses $\{Y^{(1)}, \cdots, Y^{(K)}\}$ generated from each source sequence X. Lower pairwise metric indicates high diversity between generated hypotheses.

5.5 Experimental results

5.5.1 *PermGen* v.s. Transformers

As shown in Table 2, *PermGen* can improve both the diversity and the accuracy of generated text when initialized with either non-pretrained (Transformer) or pre-trained (BART) Transformers. For example, compared with BART which has the best performance among baselines, *PermGen* reduced Self-BLEU-4 by 43.2% and improved BLEU-4 by +1.5% on AGENDA. And we observe similar improvement on all other paragraph generation tasks.

POINTER achieves the lowest performance in paragraph generation tasks. This is because its insertion operation ignores dependency between generated words so it cannot well capture the inter-sentence coherence during long-text generation.

It should be noted that since BART performed the best among all baseline methods, we apply *PermGen* on BART in the following evaluations.

5.5.2 *PermGen* v.s. Decoding Methods

We investigate the quality of text generated by *PermGen* (built on BART) when coupled with beam search, truncated sampling and nucleus sampling. Figure 4 shows that on average, *PermGen* can significantly boost diversity by 5.81% in Self-BLEU-3 and 6.83% in Self-BLEU-4, respectively, and improve accuracy by +1.2% and +1.5% in terms of Top1-BLEU-4 and Oracle-BLEU-4.

As the diversity of generated text depends on the number of produced candidates, we compare the diversity of generation between BART and *PermGen* with various number of output candidates, K. Figure 5 shows that as K increases, *PermGen* can consistently generate more diverse content, measured by the ratio of distinct 2-grams, Distinct-2 (dashed line). Meanwhile, measured by Entropy-4 (solid line), the proportion of novel words in generated candidates from *PermGen* is rising as K increases, while BART shows a flat or even falling trend.

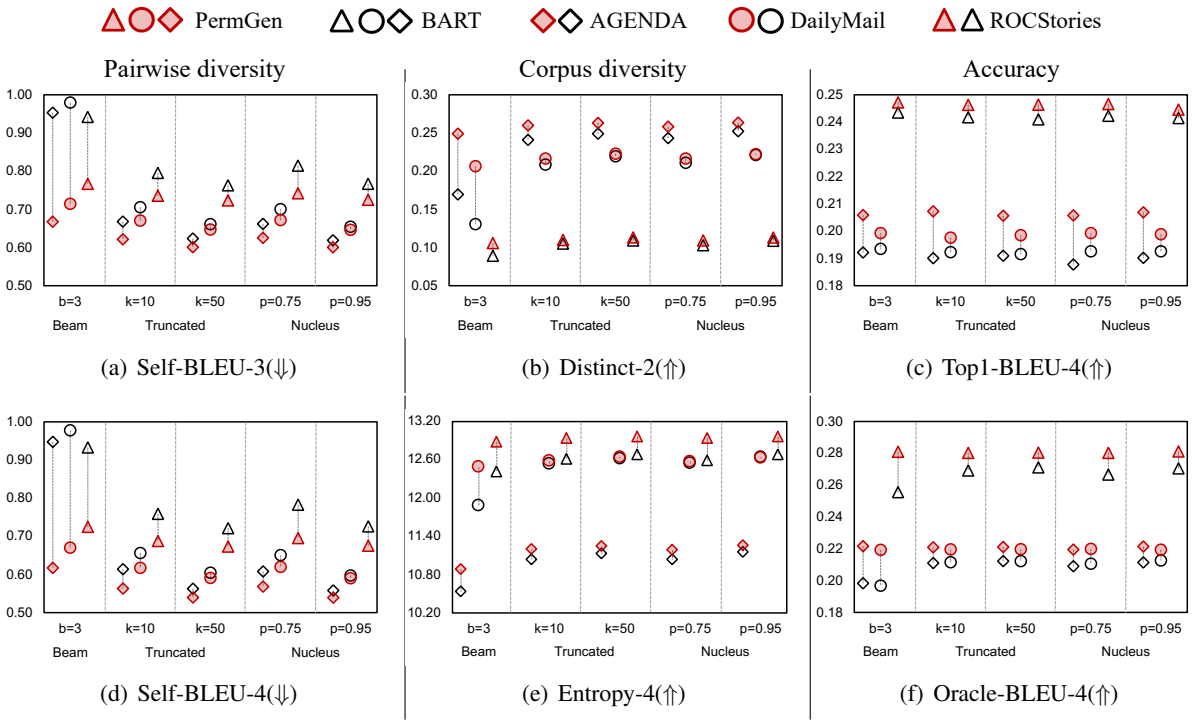

Figure 4: *PermGen* demonstrates superior performance on both diversity and accuracy compared with different diversity-promoting methods. The specific values involved in the figure are shown in Table 10 in Appendix.

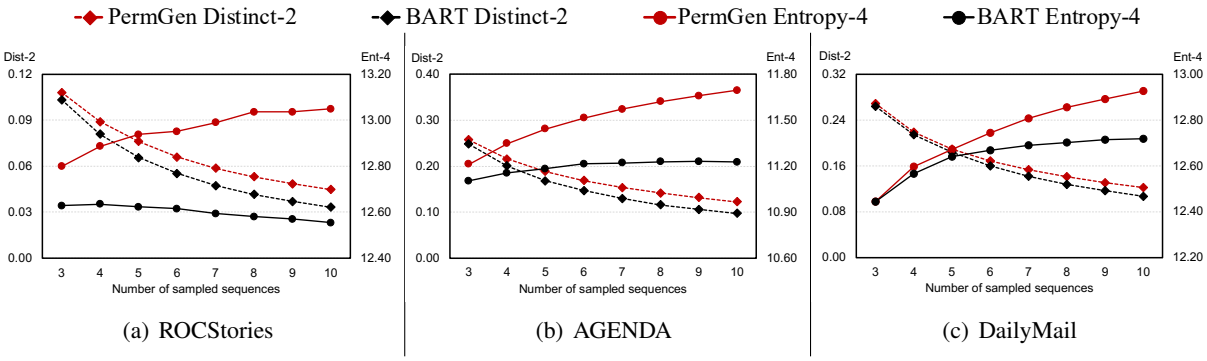

Figure 5: *PermGen* generates more diverse paragraphs over different number of sampled candidates. The diversity measured at each point is the mean value of Dist-2 and Ent-4 when $k = 10$, $k = 50$, $p = .75$, and $p = .95$.

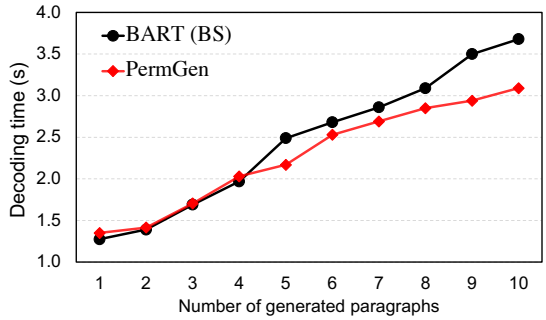

Figure 6: Time Efficiency. *PermGen* enjoys faster decoding efficiency than BART (with beam search) when generating multiple output paragraphs.

5.5.3 Human Evaluations

We sample 100 inputs from ROCStories test set and each evaluated method generates top-3 stories. Every story is assigned to five annotators with NLP background. For diversity, the annotators are given two sets of top-3 stories from two methods each time and instructed to pick the set that is more diverse. The choices are "win," "lose," or "tie." Then, the annotators give an accuracy score from 1 to 5 to measure semantic similarity between the top-1 generated story and ground truth story. Finally, the annotators need to give a fluency and coherency score from 1 to 5 for each generated story.

Table 4-5 demonstrate that *PermGen* outperforms beam search in both accuracy and fluency, while significantly improving generation diversity compared with other diversity-promoting methods.

Table 3: Case study. *PermGen* produces more diverse stories than beam search and nucleus sampling. We shade parts of the generated text which are distinct from other candidates. We provide more case studies in Appendix.

• **Inputs:** (Title) Mounting popularity ; (Keywords) started, company, friends, hard, year, slogging, reward, traction, excited
• **Beam search-1:** I started a new company with some friends . It was hard at first . After a year of slogging , I got a reward . The reward was a lot of traction . Now we are all excited to start working together . • **Beam search-2:** I started a new company with some friends . It was hard at first . After a year of slogging , I got a reward . The reward was a lot of traction . Now we are all excited . • **Beam search-3:** I started a new company with some friends . It was hard at first . After a year of slogging , I got a reward . The reward was a lot of traction . We are excited to keep doing this .
• **Nucleus sampling-1:** I started a new company with some friends . It was hard at first . After a year of slogging , the reward was a lot of traction . Now we are doing really well . I am excited to start working with my friends . • **Nucleus sampling-2:** I started a new company with some friends . It was hard at first . After a year of slogging , I got a lot of reward . The reward was a lot of traction . Now we are all excited to start working together . • **Nucleus sampling-3:** I started a new company with my friends . It was hard at first . After a year of slogging , we got a lot of reward . We got traction and are doing really well . We are excited to keep doing this .
• **PermGen-1 (reordered from [2, 1, 3, 5, 4]*):** I started a new company with some friends . I tried really hard for almost a year . It took a lot of slogging , but as a reward I got traction . Now we are all doing it together . I 'm excited to be doing it again . • **PermGen-2 (reordered from [3, 1, 2, 5, 4]):** I started a new company with some friends . It 's been hard . I 've been slogging through it as a reward for getting traction . My friends are really excited . I 'm excited to see what it 's all about . • **PermGen-3 (reordered from [5, 1, 2, 3, 4]):** I started a new company with some of my friends . It was hard at first . After a year of slogging , I got a lot of reward . I have many traction on social media . I am excited to start working with my friends .

* "Reordered from [2, 1, 3, 5, 4]" means that PermGen first generates the 2nd sentence, and then generates the 1st sentence, and so on. Finally, we reorder the generated story according to the ascending order of sentence index as shown in Figure 3.

Table 4: Human Evaluations on ROCStories: *PermGen* (ours) v.s. three baseline methods based on *diversity*.

	Win	Lose	Tie
PermGen v.s. Beam	**64.00%** (±12.71%)	14.00% (±7.70%)	22.00% (±10.73%)
PermGen v.s. Truncated	**54.80%** (±4.10%)	8.80% (±5.31%)	36.40% (±5.43%)
PermGen v.s. Nucleus	**56.00%** (±8.67%)	11.60% (±4.27%)	32.40% (±5.57%)

Table 5: Human Evaluations of *PermGen* and BART on ROCStories. **Decoding algorithm is beam search.** Minimum score is 1.0, and maximum score is 5.0.

	Accuracy	Fluency	Coherency
BART	3.34	3.93	3.85
PermGen	**3.42**	**3.97**	**3.88**

5.5.4 Time Efficiency

We vary the number of generated paragraphs to examine the decoding time of BART (with beam search) and *PermGen*. We calculate the average decoding time of a batch whose size is set as 16. The two models are run on a 32GB memory Tesla V100 GPU. As shown in Figure 6, *PermGen* enjoys faster decoding efficiency than BART when generating multiple paragraphs. This is because when generating multiple paragraphs, beam search has to spend time on ranking the generated candidates at each decoding step. However, in *PermGen*, the generation processes for different sentence orders are independent.

5.5.5 Case Study

Table 3 demonstrates generated stories from different diversity-promoting methods, including beam search, nucleus sampling and our *PermGen*. Overall, we observe that *PermGen* can generate more diverse stories than the other two methods. We notice that stories generated by beam search often differ only by punctuation and minor morphological variations, and typically only the last sentence (or last several words) is different from others. Nucleus sampling achieves better diversity than beam search, but the stories are still following similar storylines. In comparison, *PermGen* can generate semantically richer and more diverse contents.

6 Conclusions

In this paper, we proposed a novel sentence-permuted paragraph generation model, *PermGen*. *PermGen* maximizes the expected log likelihood of output paragraph *w.r.t.* all possible sentence orders. Experiments on three paragraph generation tasks demonstrated that *PermGen* outperformed original Transformer by generating more accurate and diverse text. The result is consistent on various Transformer models and decoding methods.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. *International Conference for Learning Representation (ICLR)*.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. 2019. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *International Conference for Learning Representation (ICLR)*.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Marie-Anne Lachaux, Armand Joulin, and Guillaume Lample. 2020. Target conditioning for one-to-many generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2853–2862.

Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation (COLING)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ryan L Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. 2019. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. *International Conference for Learning Representation (ICLR)*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning (ICML)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Exploring diverse expressions for paraphrase generation. 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)*, pages 3164–3173.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Yi Ren, Jinglin Liu, Xu Tan, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. A study of non-autoregressive model for sequence generation. *Proceedings of the 58th annual meeting of the Association for Computational Linguistics (ACL)*.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *Mathematical Statistics*.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics (TACL)*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International Conference on Machine Learning (ICML)*.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2018. Toward diverse text generation with inverse reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4361–4367.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *International Conference on Machine Learning (ICML)*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of conference on computer vision and pattern recognition (CVPR)*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *AAAI Conference on Artificial Intelligence*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. In *57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (EMNLP)*.

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems (NeurIPS)*.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020. Pointer: Constrained text generation via insertion-based generative pre-training. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*.

Liang Zhao, Jingjing Xu, Junyang Lin, Yichang Zhang, Hongxia Yang, and Xu Sun. 2020. Graph-based multi-hop reasoning for long text generation. *arXiv preprint arXiv:2009.13282*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*.