HIERARCHICAL SUMMARY-TO-ARTICLE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we explore *summary-to-article generation*: the task of generating long articles given a short summary, which provides finer-grained content control for the generated text. To prevent sequence-to-sequence (seq2seq) models from degenerating into language models and better controlling the long text to be generated, we propose a hierarchical generation approach which first generates a sketch of intermediate length based on the summary and then completes the article by enriching the generated sketch. To mitigate the discrepancy between the "oracle" sketch used during training and the noisy sketch generated during inference, we propose an end-to-end joint training framework based on multi-agent reinforcement learning. For evaluation, we use text summarization corpora by reversing their inputs and outputs, and introduce a novel evaluation method that employs a summarization system to summarize the generated article and test its match with the original input summary. Experiments show that our proposed hierarchical generation approach can generate a coherent and relevant article based on the given summary, yielding significant improvements upon conventional seq2seq models.

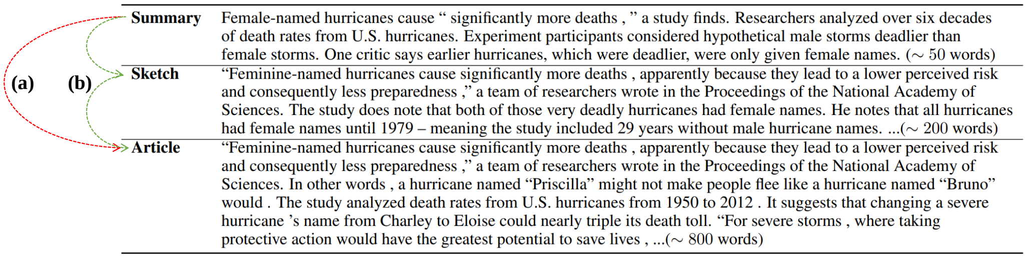
1 INTRODUCTION

In contrast to the well-explored text generation tasks like machine translation (Bahdanau et al., 2014) and text summarization (See et al., 2017), open-ended long text generation is much less explored. The existing studies on long text generation either generate long text unconditionally, such as GPT-2 (Radford et al.), or generate long text conditioning on a single sentence prompt (Fan et al., 2018; Keskar et al., 2019; Zellers et al., 2019). Although they can generate seemingly fluent text in a general domain/topic, they suffer from a lack of fine-grained control of content to be generated, which may result in generating much undesirable text and make them difficult to use in practice.

In this paper, we study long text generation with fine-grained content control. We explore *summary-to-article generation*: the task of generating a coherent and relevant long article based on a short summary of 3 to 5 sentences which summarizes the main content of the article to be generated. Compared to the previously studied unconditional or prompt-based long text generation, *summary-to-article generation* specifies the content to be generated more clearly, leading to finer-grained control of text generation. As prior work (Fan et al., 2018) points out, however, it remains challenging to generate a long, coherent and relevant article based on a summary because complex and underspecified dependencies between the summary and the article are much harder to model than the closer dependencies required for language modeling, which makes standard seq2seq models prone to degenerating into language models, neglecting salient information provided in the summary and resulting in undesirable outputs.

To address this challenge, inspired by previous work that attempts to generate text with multiple steps (Dalianis & Hovy, 1993; Reiter & Dale, 2000), we propose a hierarchical summary-to-article generation approach which decomposes the task into two subtasks: summary-to-sketch generation and sketch-to-article generation. As illustrated in Figure 1, the sketch is of an intermediate length and extracted from the article with the guidance from the summary. It serves as a draft of the output article to be generated and outlines its main content, which resembles how people plan to write long articles in their mind.

This hierarchical generation approach avoids the need for seq2seq text generation models to extend the length of source text too much, thus alleviating the aforementioned degenerating problem and enhancing the coherence and relevance of the generated text. To bridge the gap between training and inference, which arises from the discrepancy between the extracted "oracle sketch" used during



Summary	Female-named hurricanes cause " significantly more deaths , " a study finds. Researchers analyzed over six decades of death rates from U.S. hurricanes. Experiment participants considered hypothetical male storms deadlier than female storms. One critic says earlier hurricanes, which were deadlier, were only given female names. (∼ 50 words)
Sketch	"Feminine-named hurricanes cause significantly more deaths , apparently because they lead to a lower perceived risk and consequently less preparedness ," a team of researchers wrote in the Proceedings of the National Academy of Sciences. The study does note that both of those very deadly hurricanes had female names. He notes that all hurricanes had female names until 1979 – meaning the study included 29 years without male hurricane names. ...(∼ 200 words)
Article	"Feminine-named hurricanes cause significantly more deaths , apparently because they lead to a lower perceived risk and consequently less preparedness ," a team of researchers wrote in the Proceedings of the National Academy of Sciences. In other words , a hurricane named "Priscilla" might not make people flee like a hurricane named "Bruno" would . The study analyzed death rates from U.S. hurricanes from 1950 to 2012 . It suggests that changing a severe hurricane 's name from Charley to Eloise could nearly triple its death toll. "For severe storms , where taking protective action would have the greatest potential to save lives , ...(∼ 800 words)

Figure 1: Illustration of the proposed hierarchical generation model for summary-to-article generation: (a) Conventional seq2seq model generates an article directly based on the summary; (b) Our proposed hierarchical summary-to-article generation approach.

training and the noisy sketch generated during inference, we propose a gated model fusion mechanism which establishes a skip-connection from the input summary to the sketch-to-article generation model, and jointly train the summary-to-article and the sketch-to-article generation model to communicate and cooperate with each other in an end-to-end fashion with multi-agent reinforcement learning.

For evaluation, we use the text summarization corpora by reversing their inputs and outputs as our dataset. We also introduce a novel evaluation metric – ROUGE-rec which calculates how much the original summary can be reconstructed from the generated article. Experiments on the CNN/DM and BIGPATENT datasets demonstrate our proposed hierarchical generation approach can generate fluent, coherent and relevant articles based on a given summary, yielding better results than conventional seq2seq models.

Our contributions are threefold:

- We explore the task of summary-to-article generation which provides finer-grained control of generated long text and propose a hierarchical summary-to-article generation approach.
- We propose a gated model fusion mechanism and a multi-agent reinforcement learning with denoising seq2seq pretraining objective to help bridge the gap between training and inference of the hierarchical generation model.
- We propose a novel evaluation method for summary-to-article generation. Experimental results demonstrate that the proposed evaluation metric correlates better with human evaluation than the traditional metrics like perplexity for this task.

2 SUMMARY-TO-ARTICLE GENERATION

2.1 TASK DESCRIPTION

The task of summary-to-article generation aims to generate long articles ($\geq$ 500 words) based on a short summary ($\sim$ 50 words) containing several sentences which specifies the main content of the article. Given an input summary S, the output article A is expected to be fluent, coherent, relevant and faithful with respect to S. Compared to the previously studied unconditional (Radford et al.) or prompt-based (Fan et al., 2018; Zellers et al., 2019; Keskar et al., 2019) long text generation, longer summary used in our task specifies the content to be generated more clearly, leading to finer-grained control of text generation and improves the coherence of the generated article by providing richer information. This setting is hopefully more practical in real-world application scenarios, such as generating a news article from a shortlist of key fact statements, generating a patent claim from a short description, writing a story from an outline, writing a scientific paper from an abstract, etc.

2.2 HIERARCHICAL GENERATION WITH MODEL FUSION

The main challenge of generating a long article based on a summary is the information gap between the input and output as the output is much longer than the input and contains additional information. That requires the model to capture underspecified mapping from the summary to the article, which is much more difficult than the closer dependencies required for language modeling. As a consequence, the larger expansion ratio between the input and the output, the more likely the standard seq2seq

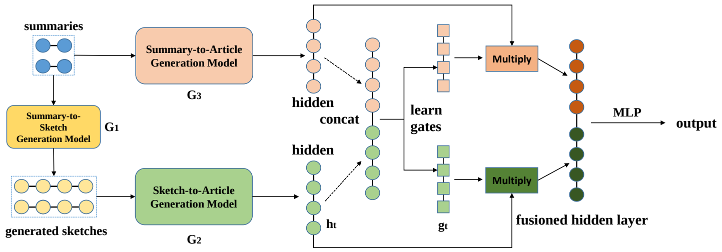


Figure 2: Overview of the proposed hierarchical summary-to-article generation model.

models are to degenerate into language models, failing to focus on the salient information provided in the summary while generating text.

Motivated by this observation, we propose a hierarchical summary-to-article generation approach which decomposes the task into two subtasks: summary-to-sketch generation and sketch-to-article generation. The sketch (we denote K in the following parts of this paper) is a draft of the final article to be generated, which outlines the main structure and the content of the article with an intermediate length. With the help of the sketch as a bridge between the short summary and the long article, hierarchical generation divides the challenge into two simpler sub-tasks, thus alleviating the issue of degeneration, facilitating the generation and enhancing the coherence and relevance of the generated article.

The overview of our proposed model is shown in Figure 2. The proposed hierarchical summary-to-article generation model mainly includes two seq2seq components: the summary-to-sketch generation model (G_1), which generates a sketch based on the summary, and the sketch-to-article generation model (G_2) which completes the article based on the generated sketch. The input summary will be first fed into the G_1 to obtain the sketch which will be then taken as input by G_2. To avoid the cases where the generated sketch is not clean and not good enough for generating articles, we add a skip-connection that is implemented as a summary-to-article generation model (G_3) to the hidden outputs of G_2's decoder, as shown in Figure 2. In this way, the hierarchical outputs of G_2 are fused with the outputs of the skip-connection from the original summary, allowing the model to adaptively learn to generate the article based on both the information from the original summary which is limited but clean and that from the generated sketch which is the more adequate but potentially noisy.

Following previous work (Fan et al., 2018), we formulate the gated model fusion mechanism as:

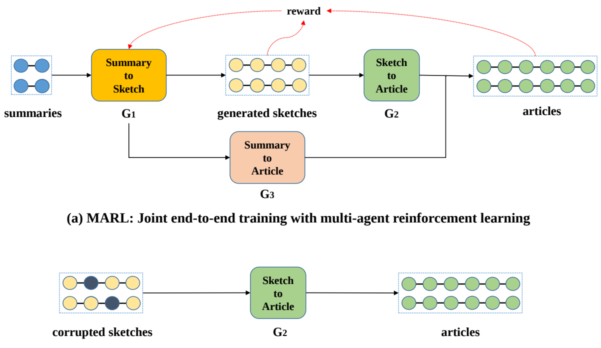
$$\begin{aligned} g_t &= \sigma\left(W\left[h_t^{\text{Sketch-Article}}; h_t^{\text{Summary-Article}}\right] + b\right) \\ h_t &= g_t \circ \left[h_t^{\text{Sketch-Article}}; h_t^{\text{Summary-Article}}\right] \end{aligned} \tag{1}$$

where $\circ$ denotes element-wise multiplication and $\sigma(\cdot)$ denotes the sigmoid function. For model fusion, the t-th decoder hidden state of the summary-to-article generation model and the sketch-to-article generation model (represented by h_t) are concatenated to learn gates g_t. The gated hidden layers are then combined by concatenation and followed by fully connected layers to generate the output token. The hierarchical generation approach with the proposed fusion mechanism is illustrated in Figure 2.

2.3 Multi-agent reinforcement learning with denoising seq2seq pretraining

The core part of the proposed decomposition approach is constructing appropriate sketches for training. Inspired by approaches for constructing supervision labels for training extractive summarization models (Nallapati et al., 2017), we propose a heuristic approach to extract important sentences in the article which are the most relevant to the summary as sketches. We compute the relevance score of each sentence a_i in the article by computing the maximum of its cosine embedding similarity with sentences s_j in the summary S under a pretrained language model[1]. Afterward, we iteratively extract the sentence of maximum relevance score from each paragraph in the article (without putting

[1] We used BERT (Devlin et al., 2018) model in our experiments as the sentence encoder.



(a) MARL: Joint end-to-end training with multi-agent reinforcement learning

(b) denoising seq2seq pretraining

Figure 3: Illustration of the proposed training strategy: (a) MARL: An end-to-end joint training framework for hierarchical summary-to-article generation; (b) denoising seq2seq pre-training for the sketch-to-article model G_2 to improve the robustness of G_2 against the noise in the generated sketch. G_1 and G_2 are updated during MARL while G_3 is fixed.

back) until the length of extracted sentences exceeds the threshold, which is empirically set to be the geometric mean of the length of the summary and the article measured by the number of tokens. This ensures the ratio of sequence length between the input and the output of the two components of our model to be roughly the same. With extracted sketches as weakly supervision, we can train the two components in our model separately with MLE.

The major limitation of training the two generation models in the proposed hierarchical generation approach separately with MLE is the discrepancy between the sketches used for training and inference. During training, the sketch used to generate the article is the "oracle" sketch that consists of sentences extracted from the article; while during inference, the sketch is generated from the summary by the summary-to-sketch model. In other words, the sketch-to-article generation model is trained to generate articles based on extracted sketches which are clean and of high quality, but receives generated sketches which are generally noisy and less adequate during inference. As a result, the gap between training and inference makes it difficult for the sketch-to-article model to work well and generate good articles in practice where no extracted sketch is available.

To address this problem and help the sketch-to-article generation model be better adapted to the generated noisy sketches during inference, we propose an end-to-end joint training framework with multi-agent reinforcement learning (MARL) and denoising pretraining to train our model, as illustrated in Figure 3.

Inspired by previous works on multi-agent communication tasks (Lowe et al., 2017; Lee et al., 2019), we model the summary-to-article generation task as a two-agent cooperation task and jointly train the agents to communicate and cooperate with each other in an end-to-end fashion with reinforcement learning. The first agent G_1 (i.e. summary-to-sketch generation model) receives a summary S as input and generates a sketch $\overline{K}$ as output message. The second agent G_2 (i.e. fusion of the sketch-to-article and the summary-to-articel generation model) is then trained to maximize the log-likelihood of the ground truth article A given the sketch message, i.e. $\log p(A|\overline{K})$. Agent G_1 is trained using REINFORCE (Williams, 1992) with reward $R = \log p_{G_2}(A|\overline{K})$. Following Lee et al. (2019), we formulate the learning objective of Agent G_1 as:

$$\mathbb{L}_{G_1} = \alpha_{\text{pg}} \left(R - \overline{R}_t\right) \log p(\overline{\mathrm{K}_t}|\overline{\mathrm{K}_{<t}}, \mathrm{S}) + \alpha_{\text{entr}} H(p(\overline{\mathrm{K}_t}|\overline{\mathrm{K}_t}, \mathrm{S})) - \alpha_{\text{b}}\, \text{MSE}\left(R, \overline{R}_t\right) \tag{2}$$

where α_{pg}, α_{entr}, α_{b} are hyperparameters, H and MSE denote entropy and mean squared error losses, $\overline{R}_t$ is a state-dependent baseline for reducing variance. The first term is the reward we aim to maximize, the second term is an entropy regularization on Agent G_1's decoder to encourage exploration, the last term is the training objective of the baseline $\overline{R}_t$.

The training objective encourages Agent G_1 to develop helpful communication policies for Agent G_2 and to generate better sketch in terms of its usefulness for Agent G_2, while also allows Agent G_2 to be adapted to noisy sketches generated by Agent G_1.

To provide a good initialization for reinforcement learning based joint training, we pretrain the summary-to-sketch generation model G_1 and the summary-to-article generation model G_3 with MLE. For the sketch-to-article generation model G_2 which suffers from the aforementioned problem of input discrepancy, we propose a denoising seq2seq pretraining objective to improve the robustness of the model with respect to the noise in the input. Specifically, we corrupt the input sketches with both word-level and sentence-level perturbation[2] and train the model to generate the same articles based on the perturbed sketches. The perturbation is expected to resemble the noise in generated sketches, thus helping the model be better adapted to the generated sketches during training.

2.4 Inference

During inference, the input summary S is fed into the summary-to-sketch generation model G_1 to generate a sketch $\overline{K}$. The generated sketch $\overline{K}$ and the original summary S is then fed to the fused sketch-to-article generation model G_2 to generate the output article. We employ the top-p (p = 0.95 in our experiments) sampling approach (Holtzman et al., 2019), which samples from the top p portion of the probability mass, expanding and contracting the candidate pool dynamically, instead of standard beam search, to avoid repetition and encourage generating diverse articles.

2.5 Evaluation

Text generation models are usually evaluated using the word-overlap based metrics (e.g., BLEU and ROUGE) and perplexity. As suggested by Liu et al. (2016), however, these metrics tend to perform poorly when evaluating open-domain text generation systems. The long and diverse nature of articles makes them even worse for evaluating the quality of summary-to-article generation models. Moreover, they are incapable of evaluating the relevance between the generated articles and the summary, which is important for evaluating the model's ability for fine-grained content control.

For better evaluating the summary-to-article text generation task, we introduce a novel evaluation metric ROUGE-rec. ROUGE-rec evaluates how much the original summary can be reconstructed from the generated article. Intuitively, if the generated article can be summarized back to the original summary (ROUGE-rec is high), that indicates that the article well focuses on the summary; on the contrary, if the generated article's summary is dissimilar to the original summary (ROUGE-rec is low), the article probably deviates from the ideas of the original summary, which is undesirable.

Formally, we define ROUGE-rec to be the ROUGE-L (Lin, 2004) score of the reconstructed summary from the generated article against the original summary. To compute ROUGE-rec, we first use a state-of-the-art abstractive summarization model (Wu et al., 2019) to summarize the generated article to obtain the reconstructed summary, and then derive the score ROUGE-rec by computing the ROUGE-L of the reconstructed summary.

3 Experiments

3.1 Datasets

We conduct experiments on the summary-to-article generation task by using text summarization datasets in the reverse direction, i.e. taking the summary as inputs and the corresponding articles as outputs. Specifically, we evaluate the proposed approaches and several baseline models on the CNN / Daily Mail (Hermann et al., 2015) dataset and the BigPatent (Sharma et al., 2019) dataset. We follow the default train-dev-test split of the original datasets. The statistics of the datasets are shown in Table 1.

Table 1: Description of the datasets used for experiments

Dataset	# Articles	Expansion Ratio	Summary Length	Article Length
CNN/DM	312,085	13.0	55.6	789.9
BigPatent	1,341,362	36.4	116.5	3572.8

[2] We present details for human evaluation in the Appendix C

Table 2: Performance on automated metrics of different models. For perplexity, the lower, the better. For dual evaluation score, the higher, the better.

Method	CNN / DM		BIGPATENT	
	PPL(gpt-2)	ROUGE-rec	PPL(gpt-2)	ROUGE-rec
seq2seq-lstm (2 layers; 512 hidden units)	36.3	11.8	17.9	11.0
seq2seq-conv	34.2	14.2	14.2	14.5
hierarchical w/o fusion	32.1	22.6	11.6	18.2
hierarchical + fusion (ours)	**30.4**	**26.5**	**11.3**	**22.6**

3.2 EVALUATION METRICS

The evaluation of open domain long text generation systems is an open problem. In our experiments, we employ the following automated metrics for evaluating the performance of compared models.

- **PPL(gpt-2)**: The perplexity of generated articles under GPT-2 (Radford et al.), one of the most powerful pretrained language model with 340M parameters, which is able to measure long range dependency. It can measure the fluency and the coherence of generated articles well.
- **ROUGE-rec**: Our proposed new metric for summary-to-article generation, which reflects how well the generated article expands the input summary, as introduced in Section 2.5.

3.3 MODEL CONFIGURATION

Following Fan et al. (2018), the basic structure of all the summary-to-sketch (G_1), sketch-to-article (G_2) and summary-to-article (G_3) models is a seq2seq convolutional model with 3-layer encoder blocks with hidden unit sizes $128 \times 2, 512$ and conlutional kernel widths 3×3, and 8-layer decoder blocks with hidden unit sizes $512 \times 4, 768 \times 2, 1024$ with convolutional kernel widths 4×8. The size of both input embedding and output embedding is 256, and the number of heads for self-attention in the decoder is 4. We keep the words which appear more than 10 times in the corpora, resulting in a vocabulary[3] of 142,971 in CNN/Daily, and 195,401 in BIGPATENT datasets. We tune the hyperparameters in Eq (2) on the validation set: $\alpha_{pg} = 1, \alpha_{entr} = 0.001, \alpha_b = 0.01$. We train the model on 4 GPUs with learning rate 0.25, dropout 0.3 and a batch of 4,000 tokens per GPU, and the pre-training of G_1, G_2 and G_3 uses the same learning configuration. We train the model with 10,000 updates for reinforcement learning and choose the best checkpoint based on the perplexity on the validation set for pretraining.

3.4 EXPERIMENTAL RESULTS

In our experiments, we compare the proposed hierarchical summary-to-article generation model with conventional LSTM seq2seq models and the convolutional seq2seq model (Fan et al., 2018) which is in the same structure with our seq2seq models.

As Table 2 shows, the perplexity of the articles generated by our proposed hierarchical model is better than that generated by the conventional seq2seq models, indicating that the hierarchical models can generate more fluent and coherent articles. Moreover, we observe the hierarchical models largely outperform the conventional seq2seq models in terms of ROUGE-rec, demonstrating their powerful capability for content control to generate articles that are more relevant and faithful to the input summary. Within the hierarchical models, we find the proposed gated model fusion mechanism plays an important role in improving the performance, because it allows the model to adaptively focus on the information from the original summary or that from the generated sketch, and also facilitates the training of the sketch-to-summary generation model by encouraging it to focus on what the summary-to-article generation model fails to learn.

To better analyze the performance of the hierarchical model, we compare the perplexity of ground-truth articles given an input summary by the conventional seq2seq convolutional model, to the perplexity of ground-truth articles given the generated sketch by the hierarchical model. According

[3] The vocabulary is shared by the encoder and decoder.

Table 3: Perplexity of ground-truth articles. PPL($S{\rightarrow}A$) denotes the perplexity by the conventional seq2seq model with input summary. PPL($K{\rightarrow}A$) denotes the perplexity by the hierarchical model with generated sketch.

Method	CNN / DM		BigPatent	
	PPL($S{\rightarrow}A$)	**PPL($K{\rightarrow}A$)**	**PPL($S{\rightarrow}A$)**	**PPL($K{\rightarrow}A$)**
seq2seq-conv	29.3	-	14.2	-
ours	**-**	**21.4**	**-**	**10.8**

to Table 3, we find that the perplexity of generating ground-truth articles by the hierarchical model is much smaller than that with the seq2seq baseline, demonstrating that the generated sketch by the hierarchical model is helpful for generating the long article. With the help of the sketch, the hierarchical model can learn the content to be generated more easily, accounting for its advantage over the seq2seq baseline.

To make the evaluation more convincing, we further conduct human evaluation. Specifically, we follow Fan et al. (2018) and invite 20 graduate students with good English proficiency as human annotators and ask them to : 1) pair shuffled summaries and generated articles, and 2) choose the better article from two articles generated by the compared model and the seq2seq-conv baseline respectively[4]. The pairing accuracy measures the relevance between the generated article and the given input summary, while human preference measures the overall quality of generated articles.

The results of human evaluation are shown in Table 4. Our approach yields consistently better results upon the seq2seq baselines in terms of both pairing accuracy and human preference. Also, the effectiveness of the fusion mechanism is verified by human evaluation.

Table 4: Human evaluation results. Pairing accuracy measures the relevance of generated articles. Human preference is compared with articles generated by the seq2seq-conv baseline and measures the overall quality of generated articles. Higher is better for both metrics.

Method	CNN / DM		BigPatent	
	Pairing Accuracy	**Human Preference**	**Pairing Accuracy**	**Human Preference**
seq2seq-conv	49.7	-	59.3	-
seq2seq-lstm	45.3	38.5	55.7	34.5
hierarchical w/o fusion	65.0	59.0	71.3	63.5
hierarchical + fusion	**68.3**	**62.5**	**74.7**	**66.0**

We calculate the sample-level Pearson correlation of our proposed automated evaluation metrics (i.e., ROUGE-rec score) as well as the conventional metrics including PPL(gpt-2), BLEU, ROUGE, and the prompt relevance test (Fan et al., 2018) with human pairing accuracy and human preference. According to Table 5, we find most conventional automated metrics like BLEU and ROUGE do not correlate well with human evaluation. Among them, the automated metrics that correlate with human score best are ROUGE-rec and PPL(GPT-2). The former one evaluate the relevance of generated article while the latter one measures the fluency and coherence of the generated article. We find that human annotators prefer articles that have better control of the content, which suggests the potential of the proposed ROUGE-rec metric in evaluating conditional long text generation models models for future works.

Table 5: Sample-level pearson correlation score of different automated metrics with human evaluation.

Pearson Correlation	ROUGE-rec	PPL(gpt-2)	BLEU	ROUGE	Prompt Relevance
- with pairing accuracy	**0.48**	0.15	0.03	-0.02	0.32
- with human preference	**0.33**	0.22	0.09	0.11	-0.05

To investigate the effects of the proposed training strategies and different choices of the length of extracted sketches, we conduct an ablation study on the CNN/DM dataset and report the result

[4] We present details for human evaluation in the Appendix C

Table 6: Ablation study of training strategies and the influence of the length of sketch. PPL($K^*\rightarrow A$) is the perplexity of ground-truth articles generated based on extracted "oracle" sketches. $0.5\times$ and $2\times$ denote the length of the sketch for training, compared with that of the geometric mean of the summary length and the article length, which is used in our model.

Method	**Automated metrics**				**Human evaluation**	
	PPL($K^*\rightarrow A$)	**PPL($K\rightarrow A$)**	**PPL(gpt-2)**	**ROUGE-rec**	**Pairing accuracy**	**Human preference**
ours	13.4	**21.4**	**30.4**	**26.5**	68.3	**62.5**
			training strategies			
- MARL w/o denoising	13.1	23.8	33.4	25.4	67.3	59.5
- w/o MARL	12.9	24.9	32.9	25.1	67.0	58.0
			length of sketch			
- shorter sketch ($0.5\times$)	17.2	23.5	33.1	24.7	65.7	57.5
- longer sketch ($2\times$)	**12.5**	24.0	32.3	24.4	**65.0**	58.0

in Table 6. We find that the proposed training strategies substantially reduce the gap between the perplexity of the articles based on extracted sketches and generated sketches and improve the fluency and coherence of generated articles measured by GPT-2, confirming their effects in bridging the gap between the training and inference stages and improving the model's robustness. As for the influence of sketch length used for training, we find that both shorter sketches and longer sketches result in sub-optimal performance, suggesting that the geometric mean of summary length and article length may be a good default choice. We hypothesize that it is because this choice makes the expansion ratio of the summary-to-sketch and sketch-to-article model identical, avoiding too much uncertainty which arises from a too large expansion ratio in either component.

For qualitative comparison, we present several samples of the article and patent claim generated by our approach and baselines in Appendix B.

4 Related work

Decomposed text generation: Our work is inspired by previous research that studies decomposing text generation into several steps. In general, the previous studies focus on either statistical template-based approaches (Wahlster et al., 1993; Dalianis & Hovy, 1993) or neural text generation models (Fan et al., 2018; Xu et al., 2018). Among the neural text generation models, most of them decompose text generation by either constructing intermediary output of roughly the same length of the final output (Fan et al., 2019; Xu et al., 2018), or generating a very short "plan" in a higher level (Yao et al., 2019). They do not address the major challenge of long text generation. In contrast, we construct sketches of intermediate length, thus providing more adequate information for generation final output and reducing the difficulty of long text generation.

Long text generation: Existing studies on long text generation either generate long text by unconditionally sampling from a pretrained language model, such as GPT-2 (Radford et al.), or generate long text conditioning on a single sentence prompt (Fan et al., 2018; Keskar et al., 2019; Zellers et al., 2019). While they can generate fluent text in a general domain/topic, they suffer from a lack of fine-grained content control of the article to be generated, which may result in generating much undesirable text and make them difficult to use in practice.

5 Conclusion

We explore the task of summary-to-article generation and propose a novel hierarchical summary-to-article generation approach. The approach first drafts a sketch that outlines the article to be generated based on the summary, then generates the article based on information in the summary and the sketch. We propose an end-to-end joint training framework through the multi-agent reinforcement learning to train the hierarchical model and evaluate its performance in multiple datasets. The experimental results show that our approach can generate a coherent and relevant article based on a given summary, outperforming the conventional seq2seq models for summary-to-article generation.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Hercules Dalianis and Eduard Hovy. Aggregation in natural language generation. In *European Workshop on Trends in Natural Language Generation*, pp. 88–105. Springer, 1993.

Abhishek Das, Satwik Kottur, Jose M. F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Angela Fan, Mike Lewis, and Yann Dauphin. Strategies for structuring story generation. *arXiv preprint arXiv:1902.01109*, 2019.

Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*, 2015.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pp. 1693–1701, 2015.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909*, 2019.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6379–6390. Curran Associates, Inc., 2017.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

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

Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

Eva Sharma, Chen Li, and Lu Wang. Bigpatent: A large-scale dataset for abstractive and coherent summarization, 2019.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. Cold fusion: Training seq2seq models together with language models. *arXiv preprint arXiv:1708.06426*, 2017.

Wolfgang Wahlster, Elisabeth André, Wolfgang Finkler, Hans-Jürgen Profitlich, and Thomas Rist. Plan-based integration of natural language and graphics generation. *Artificial intelligence*, 63(1-2): 387–427, 1993.

Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. Denoising based sequence-to-sequence pre-training for text generation, 2019.

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

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. A skeleton-based model for promoting coherence among sentences in narrative story generation. *arXiv preprint arXiv:1808.06945*, 2018.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7378–7385, 2019.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*, 2019.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*, 2019.

A Detailed related work

Long text generation The existing studies on long text generation either generate long text by unconditionally sampling from a pretrained language model, such as GPT-2 (Radford et al.), or generate long text conditioning on a single sentence prompt (Fan et al., 2018; Keskar et al., 2019; Zellers et al., 2019). Specifically, Zellers et al. (2019) pretrains a conditional language model to generate fake news based on given one-sentence headline which specifies the topic of the generated news. Keskar et al. (2019) pretrains a conditional langauge model to generate contents in the domain specified by a "domain code". Although they can generate seemingly fluent text in a general domain/topic, they suffer from a lack of fine-grained control of content to be generated, which may result in generating much undesirable text and make them difficult to use in practice.

Decomposing text generation Decomposing text generation into several steps has been explored in both statistical template-based text generation approaches (Wahlster et al., 1993; Dalianis & Hovy, 1993) and neural text generation models (Fan et al., 2018; Xu et al., 2018). Strategies for decomposing long text generation have been explored by transferring sentence compression, text summarization, and keyword extraction models to build an outline for guiding long text generation models. However, these approaches are built for generating a short "pseudo-summary" unconditionally or based on a single sentence. The intermediary training data for these approaches is thus generated and noisy. In addition, previous work on decomposing text generation generally either constructs intermediary output of roughly the same length of the final output (Fan et al., 2019; Xu et al., 2018), or generates a very short "plan" in a higher level (Yao et al., 2019). As a result, these approaches do not address the major difficulty of long text generation, which is the large difference of the length of input and output text in the seq2seq model. Indeed,the hierarchical model of Xu et al. (2018) and Yao et al. (2019) is designed for generate sentences and short stories within 50 words, and the hierarchical model of Fan et al. (2018) only generate a single sentence prompt. More recently, Fan et al. (2019) propose to first generate an action plan, then generate a anonymized story and fill in the entities in the last step. Their deconposition method extend the length of sequence in the first step, thus is orthogonal to our proposed method. Our decomposition approach is different from the aforementioned approaches in two perspectives: 1) the sketches used in our work are extracted with the guidance from both the article and the summary, thus of much better quality, and 2) we construct sketches of intermediate length, thus providing more adequate information for generation final output and reducing the difficulty of expanding the length of input by a large ratio in one pass.

Denoising pretraining for seq2seq models Pretraining a denoising autoencoder for text generation is explored in recent works (Edunov et al., 2018; Lample et al., 2017; Wang et al., 2019; Zhao et al., 2019). The motivation of their approaches is to tackle the data sparsity problem while we employ the denoising objective for training the sketch-to-article generation model to be better adapted to generated sketches. As a result, the corruption methods in our work are different and our model is trained to directly output the target sequence instead of reconstructing the original input.

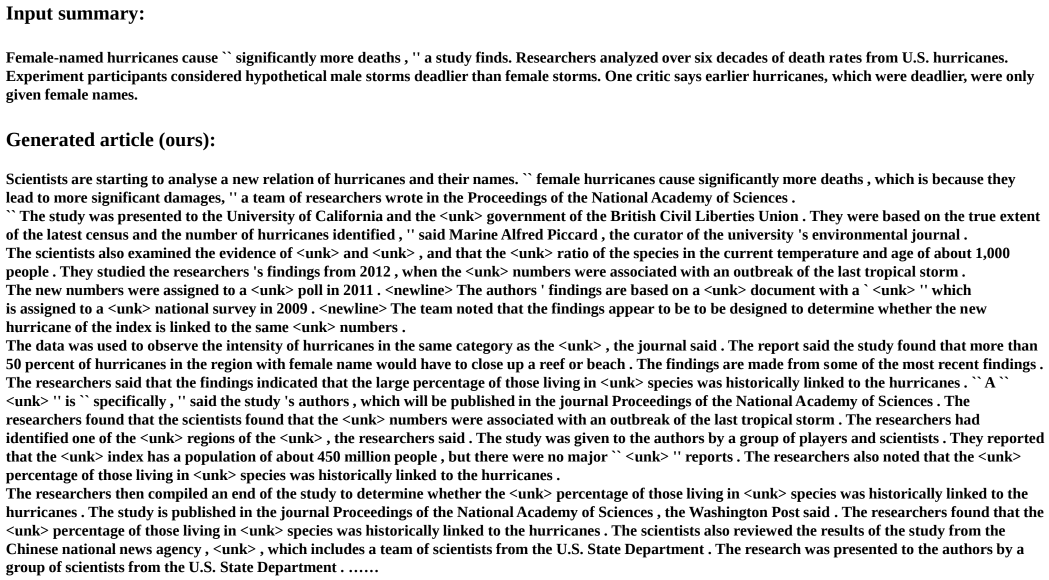
Learning to communicate and cooperate between multiple agents The idea of training multiple agents to communicate and cooperate with each other for accomplishing a common goal is well explored in multi-agent reinforcement learning literatures (Lowe et al., 2017; Foerster et al., 2018; Das et al., 2017). The most similar work to ours is that of Lee et al. (2019), which pretrains two translation models of Fr-En and En-De respectively, and train them to perform Fr-De translation cooperatively with reinforcement learning.

Model fusion Previous work has investigated the integration of language models with seq2seq models and the fusion of two identical seq2seq models. Gulcehre et al. (2015) combined a trained language model with a trained seq2seq model to learn a gating function that joins them. Sriram et al. (2017) propose training the seq2seq model given the fixed language model then learning a gate to filter the information from the language model. Fan et al. (2018) propose to train another identical seq2seq model based on a pretrained seq2seq model. To our knowledge, our work is the first to investigate the fusion of two seq2seq model with different input and the same output to combine different source of information and prevent error-propagation.

B GENERATED SAMPLES

We present samples in both datasets generated by both baseline convolutional seq2seq architecture and the proposed model in this section to facilitate qualitative analysis. We highlight repetitive parts of generated article in red.

B.1 SAMPLES ON CNN / DM DATASET



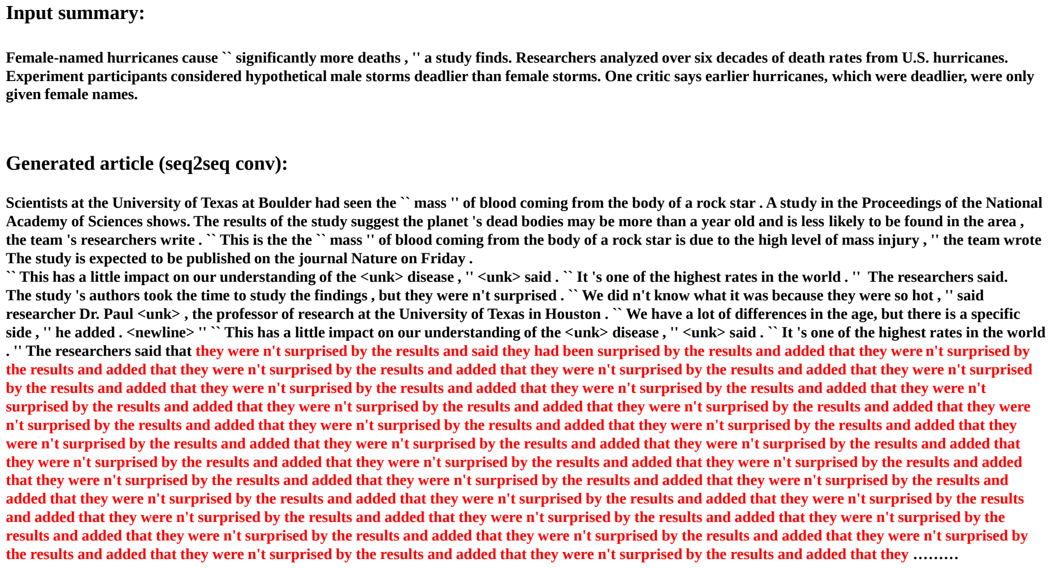
Input summary:

Female-named hurricanes cause `` significantly more deaths , '' a study finds. Researchers analyzed over six decades of death rates from U.S. hurricanes. Experiment participants considered hypothetical male storms deadlier than female storms. One critic says earlier hurricanes, which were deadlier, were only given female names.

Generated article (ours):

Scientists are starting to analyse a new relation of hurricanes and their names. `` female hurricanes cause significantly more deaths , which is because they lead to more significant damages, '' a team of researchers wrote in the Proceedings of the National Academy of Sciences .
`` The study was presented to the University of California and the <unk> government of the British Civil Liberties Union . They were based on the true extent of the latest census and the number of hurricanes identified , '' said Marine Alfred Piccard , the curator of the university 's environmental journal .
The scientists also examined the evidence of <unk> and <unk> , and that the <unk> ratio of the species in the current temperature and age of about 1,000 people . They studied the researchers 's findings from 2012 , when the <unk> numbers were associated with an outbreak of the last tropical storm .
The new numbers were assigned to a <unk> poll in 2011 . <newline> The authors ' findings are based on a <unk> document with a ` <unk> '' which is assigned to a <unk> national survey in 2009 . <newline> The team noted that the findings appear to be to be designed to determine whether the new hurricane of the index is linked to the same <unk> numbers .
The data was used to observe the intensity of hurricanes in the same category as the <unk> , the journal said . The report said the study found that more than 50 percent of hurricanes in the region with female name would have to close up a reef or beach . The findings are made from some of the most recent findings . The researchers said that the findings indicated that the large percentage of those living in <unk> species was historically linked to the hurricanes . `` A `` <unk> '' is `` specifically , '' said the study 's authors , which will be published in the journal Proceedings of the National Academy of Sciences . The researchers found that the scientists found that the <unk> numbers were associated with an outbreak of the last tropical storm . The researchers had identified one of the <unk> regions of the <unk> , the researchers said . The study was given to the authors by a group of players and scientists . They reported that the <unk> index has a population of about 450 million people , but there were no major `` <unk> '' reports . The researchers also noted that the <unk> percentage of those living in <unk> species was historically linked to the hurricanes .
The researchers then compiled an end of the study to determine whether the <unk> percentage of those living in <unk> species was historically linked to the hurricanes . The study is published in the journal Proceedings of the National Academy of Sciences , the Washington Post said . The researchers found that the <unk> percentage of those living in <unk> species was historically linked to the hurricanes . The scientists also reviewed the results of the study from the Chinese national news agency , <unk> , which includes a team of scientists from the U.S. State Department . The research was presented to the authors by a group of scientists from the U.S. State Department . ……

Figure 4: Sample generated by our model on the CNN / Daily Mail dataset.



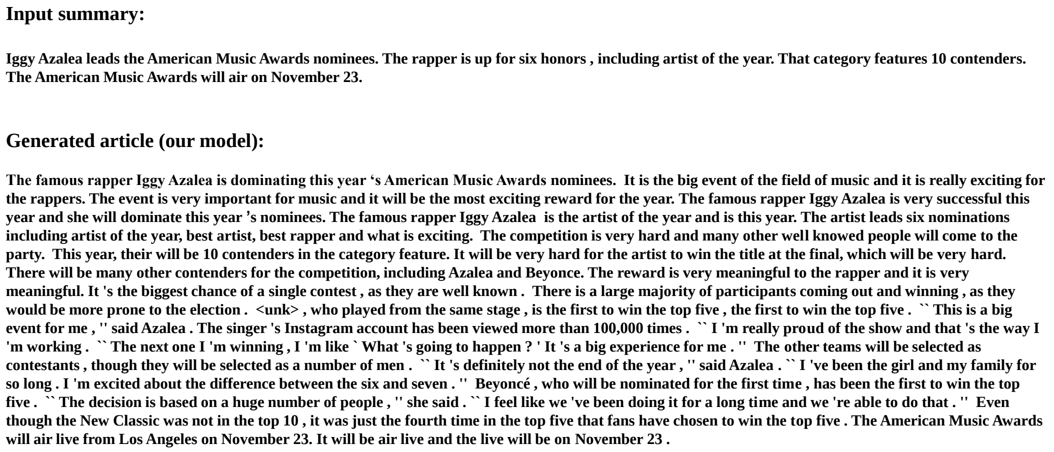
Input summary:

Female-named hurricanes cause `` significantly more deaths , '' a study finds. Researchers analyzed over six decades of death rates from U.S. hurricanes. Experiment participants considered hypothetical male storms deadlier than female storms. One critic says earlier hurricanes, which were deadlier, were only given female names.

Generated article (seq2seq conv):

Scientists at the University of Texas at Boulder had seen the `` mass '' of blood coming from the body of a rock star . A study in the Proceedings of the National Academy of Sciences shows. The results of the study suggest the planet 's dead bodies may be more than a year old and is less likely to be found in the area , the team 's researchers write . `` This is the the `` mass '' of blood coming from the body of a rock star is due to the high level of mass injury , '' the team wrote The study is expected to be published on the journal Nature on Friday .
`` This has a little impact on our understanding of the <unk> disease , '' <unk> said . `` It 's one of the highest rates in the world . '' The researchers said. The study 's authors took the time to study the findings , but they were n't surprised . `` We did n't know what it was because they were so hot , '' said researcher Dr. Paul <unk> , the professor of research at the University of Texas in Houston . `` We have a lot of differences in the age, but there is a specific side , '' he added . <newline> '' `` This has a little impact on our understanding of the <unk> disease , '' <unk> said . `` It 's one of the highest rates in the world . '' The researchers said that they were n't surprised by the results and said they had been surprised by the results and added that they were n't surprised by the results and added that they ………

Figure 5: Sample generated by baseline model on the CNN / Daily Mail dataset. We highlight repetitive parts of generated article in red.



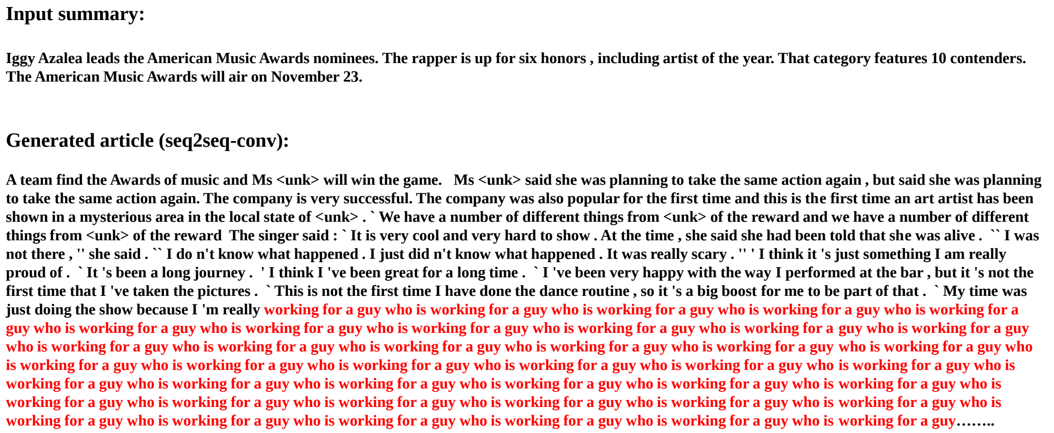
Input summary:

Iggy Azalea leads the American Music Awards nominees. The rapper is up for six honors , including artist of the year. That category features 10 contenders. The American Music Awards will air on November 23.

Generated article (our model):

The famous rapper Iggy Azalea is dominating this year ‘s American Music Awards nominees. It is the big event of the field of music and it is really exciting for the rappers. The event is very important for music and it will be the most exciting reward for the year. The famous rapper Iggy Azalea is very successful this year and she will dominate this year ’s nominees. The famous rapper Iggy Azalea is the artist of the year and is this year. The artist leads six nominations including artist of the year, best artist, best rapper and what is exciting. The competition is very hard and many other well knowed people will come to the party. This year, their will be 10 contenders in the category feature. It will be very hard for the artist to win the title at the final, which will be very hard. There will be many other contenders for the competition, including Azalea and Beyonce. The reward is very meaningful to the rapper and it is very meaningful. It 's the biggest chance of a single contest , as they are well known . There is a large majority of participants coming out and winning , as they would be more prone to the election . <unk> , who played from the same stage , is the first to win the top five , the first to win the top five . `` This is a big event for me , '' said Azalea . The singer 's Instagram account has been viewed more than 100,000 times . `` I 'm really proud of the show and that 's the way I 'm working . `` The next one I 'm winning , I 'm like ` What 's going to happen ? ' It 's a big experience for me . '' The other teams will be selected as contestants , though they will be selected as a number of men . `` It 's definitely not the end of the year , '' said Azalea . `` I 've been the girl and my family for so long . I 'm excited about the difference between the six and seven . '' Beyoncé , who will be nominated for the first time , has been the first to win the top five . `` The decision is based on a huge number of people , '' she said . `` I feel like we 've been doing it for a long time and we 're able to do that . '' Even though the New Classic was not in the top 10 , it was just the fourth time in the top five that fans have chosen to win the top five . The American Music Awards will air live from Los Angeles on November 23. It will be air live and the live will be on November 23 .

Figure 6: Sample generated by our model on the CNN / Daily Mail dataset.



Input summary:

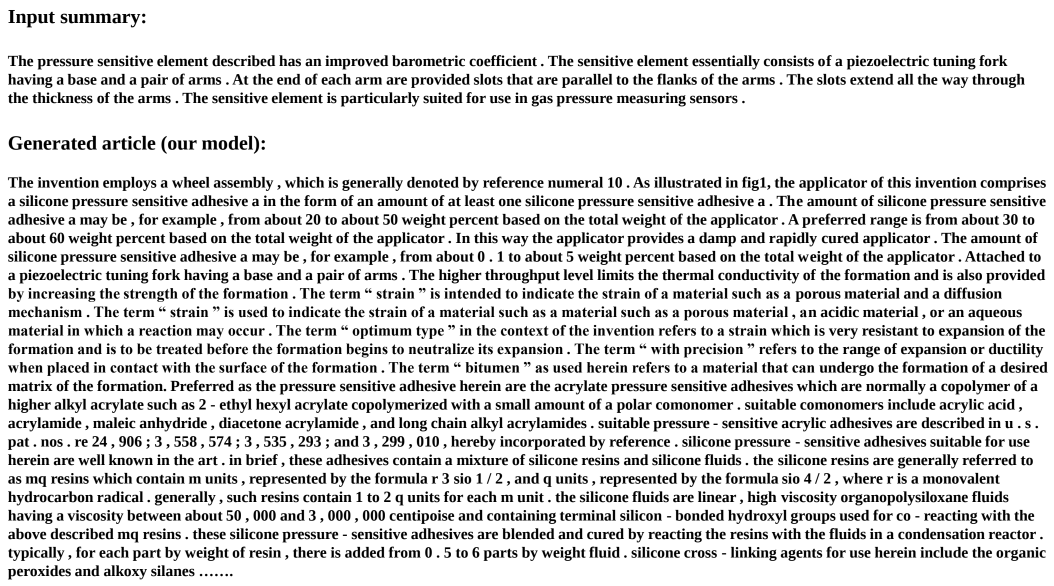
Iggy Azalea leads the American Music Awards nominees. The rapper is up for six honors , including artist of the year. That category features 10 contenders. The American Music Awards will air on November 23.

Generated article (seq2seq-conv):

A team find the Awards of music and Ms <unk> will win the game. Ms <unk> said she was planning to take the same action again , but said she was planning to take the same action again. The company is very successful. The company was also popular for the first time and this is the first time an art artist has been shown in a mysterious area in the local state of <unk> . ` We have a number of different things from <unk> of the reward and we have a number of different things from <unk> of the reward The singer said : ` It is very cool and very hard to show . At the time , she said she had been told that she was alive . `` I was not there , '' she said . `` I do n't know what happened . I just did n't know what happened . It was really scary . '' ' I think it 's just something I am really proud of . ` It 's been a long journey . ' I think I 've been great for a long time . ` I 've been very happy with the way I performed at the bar , but it 's not the first time that I 've taken the pictures . ` This is not the first time I have done the dance routine , so it 's a big boost for me to be part of that . ` My time was just doing the show because I 'm really working for a guy who is working for a guy……..

Figure 7: Sample generated by baseline model on the CNN / Daily Mail dataset. We highlight repetitive parts of generated article in red.

B.2 SAMPLES ON BIGPATENT DATASET



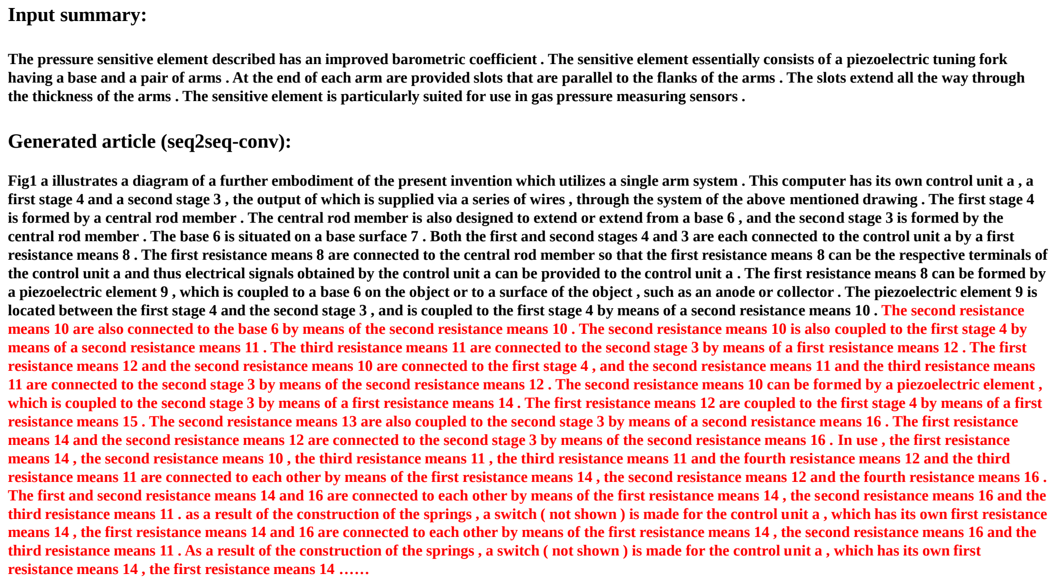
Input summary:

The pressure sensitive element described has an improved barometric coefficient . The sensitive element essentially consists of a piezoelectric tuning fork having a base and a pair of arms . At the end of each arm are provided slots that are parallel to the flanks of the arms . The slots extend all the way through the thickness of the arms . The sensitive element is particularly suited for use in gas pressure measuring sensors .

Generated article (our model):

The invention employs a wheel assembly , which is generally denoted by reference numeral 10 . As illustrated in fig1, the applicator of this invention comprises a silicone pressure sensitive adhesive a in the form of an amount of at least one silicone pressure sensitive adhesive a . The amount of silicone pressure sensitive adhesive a may be , for example , from about 20 to about 50 weight percent based on the total weight of the applicator . A preferred range is from about 30 to about 60 weight percent based on the total weight of the applicator . In this way the applicator provides a damp and rapidly cured applicator . The amount of silicone pressure sensitive adhesive a may be , for example , from about 0 . 1 to about 5 weight percent based on the total weight of the applicator . Attached to a piezoelectric tuning fork having a base and a pair of arms . The higher throughput level limits the thermal conductivity of the formation and is also provided by increasing the strength of the formation . The term “ strain ” is intended to indicate the strain of a material such as a porous material and a diffusion mechanism . The term “ strain ” is used to indicate the strain of a material such as a material such as a porous material , an acidic material , or an aqueous material in which a reaction may occur . The term “ optimum type ” in the context of the invention refers to a strain which is very resistant to expansion of the formation and is to be treated before the formation begins to neutralize its expansion . The term “ with precision ” refers to the range of expansion or ductility when placed in contact with the surface of the formation . The term “ bitumen ” as used herein refers to a material that can undergo the formation of a desired matrix of the formation. Preferred as the pressure sensitive adhesive herein are the acrylate pressure sensitive adhesives which are normally a copolymer of a higher alkyl acrylate such as 2 - ethyl hexyl acrylate copolymerized with a small amount of a polar comonomer . suitable comonomers include acrylic acid , acrylamide , maleic anhydride , diacetone acrylamide , and long chain alkyl acrylamides . suitable pressure - sensitive acrylic adhesives are described in u . s . pat . nos . re 24 , 906 ; 3 , 558 , 574 ; 3 , 535 , 293 ; and 3 , 299 , 010 , hereby incorporated by reference . silicone pressure - sensitive adhesives suitable for use herein are well known in the art . in brief , these adhesives contain a mixture of silicone resins and silicone fluids . the silicone resins are generally referred to as mq resins which contain m units , represented by the formula r 3 sio 1 / 2 , and q units , represented by the formula sio 4 / 2 , where r is a monovalent hydrocarbon radical . generally , such resins contain 1 to 2 q units for each m unit . the silicone fluids are linear , high viscosity organopolysiloxane fluids having a viscosity between about 50 , 000 and 3 , 000 , 000 centipoise and containing terminal silicon - bonded hydroxyl groups used for co - reacting with the above described mq resins . these silicone pressure - sensitive adhesives are blended and cured by reacting the resins with the fluids in a condensation reactor . typically , for each part by weight of resin , there is added from 0 . 5 to 6 parts by weight fluid . silicone cross - linking agents for use herein include the organic peroxides and alkoxy silanes …….

Figure 8: Sample generated by our model on the CNN / Daily Mail dataset.



Input summary:

The pressure sensitive element described has an improved barometric coefficient . The sensitive element essentially consists of a piezoelectric tuning fork having a base and a pair of arms . At the end of each arm are provided slots that are parallel to the flanks of the arms . The slots extend all the way through the thickness of the arms . The sensitive element is particularly suited for use in gas pressure measuring sensors .

Generated article (seq2seq-conv):

Fig1 a illustrates a diagram of a further embodiment of the present invention which utilizes a single arm system . This computer has its own control unit a , a first stage 4 and a second stage 3 , the output of which is supplied via a series of wires , through the system of the above mentioned drawing . The first stage 4 is formed by a central rod member . The central rod member is also designed to extend or extend from a base 6 , and the second stage 3 is formed by the central rod member . The base 6 is situated on a base surface 7 . Both the first and second stages 4 and 3 are each connected to the control unit a by a first resistance means 8 . The first resistance means 8 are connected to the central rod member so that the first resistance means 8 can be the respective terminals of the control unit a and thus electrical signals obtained by the control unit a can be provided to the control unit a . The first resistance means 8 can be formed by a piezoelectric element 9 , which is coupled to a base 6 on the object or to a surface of the object , such as an anode or collector . The piezoelectric element 9 is located between the first stage 4 and the second stage 3 , and is coupled to the first stage 4 by means of a second resistance means 10 . The second resistance means 10 are also connected to the base 6 by means of the second resistance means 10 . The second resistance means 10 is also coupled to the first stage 4 by means of a second resistance means 11 . The third resistance means 11 are connected to the second stage 3 by means of a first resistance means 12 . The first resistance means 12 and the second resistance means 10 are connected to the first stage 4 , and the second resistance means 11 and the third resistance means 11 are connected to the second stage 3 by means of the second resistance means 12 . The second resistance means 10 can be formed by a piezoelectric element , which is coupled to the second stage 3 by means of a first resistance means 14 . The first resistance means 12 are coupled to the first stage 4 by means of a first resistance means 15 . The second resistance means 13 are also coupled to the second stage 3 by means of a second resistance means 16 . The first resistance means 14 and the second resistance means 12 are connected to the second stage 3 by means of the second resistance means 16 . In use , the first resistance means 14 , the second resistance means 10 , the third resistance means 11 , the third resistance means 11 and the fourth resistance means 12 and the third resistance means 11 are connected to each other by means of the first resistance means 14 , the second resistance means 12 and the fourth resistance means 16 . The first and second resistance means 14 and 16 are connected to each other by means of the first resistance means 14 , the second resistance means 16 and the third resistance means 11 . as a result of the construction of the springs , a switch (not shown) is made for the control unit a , which has its own first resistance means 14 , the first resistance means 14 and 16 are connected to each other by means of the first resistance means 14 , the second resistance means 16 and the third resistance means 11 . As a result of the construction of the springs , a switch (not shown) is made for the control unit a , which has its own first resistance means 14 , the first resistance means 14 ……

Figure 9: Sample generated by baseline model on the CNN / Daily Mail dataset. We highlight repetitive parts of generated article in red.

Input summary:

A powder - spraying apparatus having simultaneous internal and external charging , which is suitable for electrostatic powder coating . In order to achieve improved constancy of the deposition efficiency , largely independent of the distance between the powder - spraying apparatus and the work piece , it is proposed to connect both internal high - voltage electrodes and external high - voltage electrodes to a high - voltage source via high - resistance components .

Generated article (our model):

The presented powder - spraying apparatus shown in fig1 comprises : a frame 1 , a mass - spraying apparatus 4 and a work piece 7 . The powder - spraying apparatus 3 has an inert gas inlet 5 for emitting powder . In the form of an anode 6 , the powder - spraying apparatus is capable of being connected to a gas source . the powder - spraying apparatus 3 includes a powder - spraying structure . For the purpose of the present invention , the powder - spraying structure comprises a particle - spraying strip 7 , for example of the form of a foil or foil - foil , which is situated in a plurality of solid filaments 8 (see fig2). The powder - spraying structure comprises an anode 9 disposed inside a plurality of solid filaments 10 (see fig2). In the form of an anode 9 , the cathode 9 is connected to a terminal 13 of the gas source . In this case , the anode 9 and the anode 9 are also connected to a terminal 14 . The motor assembly 46 has a pinion 88 , which is comprised of five teeth 90 . A clutch 65 is also attached to housing 52 , and is located on the top of clutch 65 . The clutch 65 is preferably a double - acting clutch . The alternative is to use dry process like electrostatic powder coating or other transport method by air and gas . besides , these balls can be optionally used as the mask for rie to make the gan surface rougher to enhance the extraction efficiency further . Fig3 shows a multi - layer epitaxial structure of an exemplary ingan led formed on a carrier , which can be a sapphire substrate in one embodiment . The multi - layer epitaxial structure formed above the sapphire substrate includes an n - gan based layer 42 , an mqw active layer 44 and a contact layer 46 . The n - gan based layer 42 may be a doped n - gan based layer , such as one doped with si for electricity conduction , having a thickness of 2 - 6 microns , for example . The mqw active layer 44 can be an ingan / gan mqw active layer . In order to achieve improved constancy of the deposition efficiency , largely independent of the distance between the powder - spraying apparatus and the work piece , it is proposed to connect both internal high - voltage electrodes and external high - voltage electrodes to a high - voltage source via high - resistance components . A internal high - voltage electrodes 32 are inserted into the ring 18. The ring 18 can be connected to a high - voltage 6 via a high - resistance rod 3 (see fig4). A connection point 11 for the rod 3 is located at a center of a ring portion located between two of the electrodes 32 . Moreover , various features are described which may be exhibited by some embodiments and not by others . Similarly , various requirements are described which may be requirements for some embodiments but not other embodiments . The term " component " as used herein , is intended to cover a structure including a plurality of components and components which may be included in some embodiments of the invention . Thus , the appearance of such components or other elements is not restricted by such features . The terms " component ," " component ", and " component " in terms of these terms may include any means that may be used to describe the presence of one or more other features , structures , or characteristics . The appearances of the phrase " in one embodiment " in various places in the specification are not necessarily all referring to the same embodiment ……

Figure 10: Sample generated by our model on the CNN / Daily Mail dataset.

Input summary:

A powder - spraying apparatus having simultaneous internal and external charging , which is suitable for electrostatic powder coating . In order to achieve improved constancy of the deposition efficiency , largely independent of the distance between the powder - spraying apparatus and the work piece , it is proposed to connect both internal high - voltage electrodes and external high - voltage electrodes to a high - voltage source via high - resistance components .

Generated article (seq2seq-conv):

The novel powder - spraying apparatus shown in fig1 comprises : a frame 2 , a component 3 and a work piece 4 . The powder - spraying structure comprises an anode 9 disposed inside a plurality of solid filaments 10. Each anode 9 is connected to a terminal 11 of the gas source via a terminal 14 . the terminal 11 of the gas source is connected to the powder - spraying structure . the clutch 92 engages the gear 84 , which is secured to the motor shaft 18 . a brake 46 is connected between the motor shaft 18 and the clutch 92 , which is secured to the motor shaft 18 , by means of a hinge 92 . the clutch 92 is attached to the motor shaft 18 and is positioned to engage the motor shaft 18 . the clutch 92 is adapted to engage the motor shaft 18 , with the same direction as the clutch 92 , and this direction is selected such that the clutch 92 is capable of maintaining the rotation of the motor shaft 18 . the clutch 92 is attached to the motor shaft 18 by means of a pivot pin 96 . the motor assembly 46 has an upper drive drive assembly 16 , a drive member 40 , 42 , having a drive member 22 , 24 , and a drive member 40 , 42 , which is mounted to a lower end of housing 52 , as shown in fig1 . the upper drive drive assembly 16 , and drive members 22 and 40 are attached to the housing 52 by means of an upper and lower pivots 102 and 104 , which are preferably secured to housing 52 by means of a slot 104 . the powder - spraying structure further comprises an anode 16 , the cathode 15 , the cathode 17 , the anode 16 , the cathode 17 , the anode 16 , the cathode 17 and the anode 16 . the powder - spraying structure further comprises an anode 18 , the cathode 19 , the anode 20 , the anode 21 , the anode 22 , the cathode 23 , the anode 24 , the cathode 25 , the anode 26 , the cathode 27 , the cathode 28 , the anode 29 , the anode 30 , the cathode 31 , the cathode 32 , the cathode 33 , the cathode 34 , the cathode 35 , the anode 36 , the cathode 37 , the anode 38 , the cathode 39 , the anode 40 , the cathode 41 , the cathode 42 , the cathode 43 , the anode 44 , the cathode 45 , the cathode 46 , the anode 48 , the anode 50 , the anode 51 , the anode 52 , the cathode 53 , the anode 54 , the cathode 55 , the anode 56 , the anode 57 , the anode 58 , the anode 60 , the anode 62 , the cathode 65 , the anode 68 , the anode 70 , the anode 72 , the anode 74 , the anode 76 , the anode 78 , the anode 80 , the cathode 82 , the anode 84 , the cathode 86 , the anode 88 , the anode 90 , the anode 92 , the anode 94 , the anode 102 , the cathode 104 , the anode 106 , the anode 108 , the anode 110 , the cathode 112 , the anode 114 , the cathode 116 , the anode 118 , the anode 166 , the anode 190 , the anode 170 , the cathode 182 , the anode 174 , the anode 178 , the cathode 180 , the anode 126 , the anode 128 , the cathode 182 , the anode 192 , the cathode 128 , the anode 184 , the cathode 194 , the anode 168 , the anode 184 , the cathode 175 , the anode 182 , the cathode 226 , the anode 186 , the cathode 186 , the anode 186 , the cathode 156 , the anode 158 , the cathode 158 , the anode 160, the cathode 180 , the anode 126 , the anode 128 , the cathode 182 , the anode 192 , the cathode 128 , the anode 184 , the cathode 194 , the anode 168 , the anode 184 , the cathode 175 , the anode 182 , the cathode 226 , the anode 186 , the cathode 186 , the anode 186 , the cathode 156 , the anode 158 , the cathode 158 , the anode 160 ……

Figure 11: Sample generated by baseline model on the CNN / Daily Mail dataset. We highlight repetitive parts of generated article in red.

C Detailed experimental setting

C.1 Perturbation approaches

Sentence-level Perturbations We consider the following operations 1) *Shuf* that shuffles the sequence of sentences in the extracted sketches, 2) *Drop* that completely drops certain sentences, and 4) *Repl* that randomly replace one sentence in the extracted sketches by another sentence in the dataset.

Word-level perturbations We consider similar operations but at the word level within every sentence 1) *word-shuffle* that randomly shuffles the words within a sentence 2) *reverse* that reverses the ordering of words, 3) *word-drop* that drops 30% of the words in one sentence uniformly 4) *noun-drop* that drops all nouns, 5) *verb-drop* that drops all verbs, and 6) *word-repl* that replace 30% of words with a random word in the vocabulary uniformly.

We explain the role of different perturbation and their potential effects briefly. The *Shuf* and *reverse* perturbations change the chronological order of sentences and denoising seq2seq pretraining with these kinds of perturbation may help the model to be robust when the summary-to-sketch generation model fails to generate sentences in chronological order. The *Drop* and *Repl* perturbations may help the model to be robust when some information is lost during summary-to-sketch generation.

C.2 Preprocessing details

For news articles in the CNN/DM dataset, we truncate them to be at most 1000 tokens, for patent claims, we truncate them to 3000 tokens considering the limit of GPU memory. We tokenize training data with moses tokenizer and did not use byte-pair encoding following Fan et al. (2018).

C.3 Human evaluation

For human evaluation, we invite 20 graduate students with good English proficiency as human annotators. For each dataset, we randomly sample 100 summaries from the test set and generate articles with them using each compared models and distribute them randomly to human annotators. Human annotators are required to perform two tasks: 1) the triple pairing task, where groups of three articles are presented to the human judges. The articles and their corresponding prompts are shuffled, and human annotators are asked to select the correct pairing for all three prompts. The accuracy is used to measure the relevance between generated articles and corresponding summaries. 2) human preference task, where human annotators are shown two different articles generated by a compared model and the seq2seq-conv model respectively, together with the same summary based on which the articles are generated. Annotators are then asked to mark which article they prefer. Each generated news article is paired 3 times and appears in the preference test 2 times, so that each model get 300/200 results in the triple pairing task and the human preference task respectively. Patent claims are strimmed to 200 and 400 words respectively to ease human evaluation.