Select and Attend: Towards Controllable Content Selection in Text Generation

Xiaoyu Shen[1,2]*, **Jun Suzuki**[3,4], **Kentaro Inui**[3,4], **Hui Su**[5]
Dietrich Klakow[1] and **Satoshi Sekine**[4]
[1]Spoken Language Systems (LSV), Saarland University, Germany
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3]Tohoku University [4]RIKEN AIP, Japan
[5]Pattern Recognition Center, Wechat AI, Tencent Inc, China

Abstract

Many text generation tasks naturally contain two steps: content selection and surface realization. Current neural encoder-decoder models conflate both steps into a black-box architecture. As a result, the content to be described in the text cannot be explicitly controlled. This paper tackles this problem by decoupling content selection from the decoder. The decoupled content selection is human interpretable, whose value can be manually manipulated to control the content of generated text. The model can be trained end-to-end without human annotations by maximizing a lower bound of the marginal likelihood. We further propose an effective way to trade-off between performance and controllability with a single adjustable hyperparameter. In both data-to-text and headline generation tasks, our model achieves promising results, paving the way for controllable content selection in text generation.[1]

1 Introduction

Many text generation tasks, e.g., data-to-text, summarization and image captioning, can be naturally divided into two steps: content selection and surface realization. The generations are supposed to have two levels of diversity: (1) content-level diversity reflecting multiple possibilities of content selection (what to say) and (2) surface-level diversity reflecting the linguistic variations of verbalizing the selected contents (how to say) (Reiter and Dale, 2000; Nema et al., 2017). Recent neural network models handle these tasks with the encoder-decoder (Enc-Dec) framework (Sutskever et al., 2014; Bahdanau et al., 2015), which simultaneously performs selecting and verbalizing in a black-box way. Therefore, both levels of diversity are entangled within the generation. This entanglement, however, sacrifices the controllability and interpretability, making it diffifcult to specify the content to be conveyed in the generated text (Qin et al., 2018; Wiseman et al., 2018).

Source Sentence: The sri lankan government on Wednesday announced the closure of government schools with immediate effect as a military campaign against tamil separatists escalated in the north of the country.

Selected :	sri lankan, closure, schools
Text:	sri lanka closes schools .
Selected :	sri lankan, Wednesday, closure, schools
Text:	sri lanka closes schools on Wednesday.
Selected :	sri lankan, closure, schools, military campaign
Text:	sri lanka shuts down schools amid war fears.
Selected :	sri lankan, announced, closure, schools
Text:	sri lanka declares closure of schools.

Table 1: Headline generation examples from our model. We can generate text describing various contents by sampling different content selections. The selected source word and its corresponding realizations in the text are highlighted with the same color.

With this in mind, this paper proposes decoupling content selection from the Enc-Dec framework to allow finer-grained control over the generation. Table 1 shows an example. We can easily modify the content selection to generate text with various focuses, or sample multiple paraphrases by fixing the content selection.

Though there has been much work dealing with content selection for the Enc-Dec, none of them is able to address the above concerns properly. Current methods can be categorized into the following three classes and have different limits:

1. **Bottom-up**: Train a separate content selector to constrain the attention to source tokens (Gehrmann et al., 2018), but the separate training of selector/generator might lead to

*Work mostly done while at RIKEN AIP. Correspondence to xshen@mpi-inf.mpg.de

[1]The source code is available on https://github.com/chin-gyou/controllable-selection

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 579–590,
Hong Kong, China, November 3–7, 2019.

discrepancy when integrating them together.

2. **Soft-select**: Learn a soft mask to filter useless information (Mei et al., 2016; Zhou et al., 2017). However, the mask is *deterministic* without any probabilistic variations, making it hard to model the content-level diversity.

3. **Reinforce-select**: Train the selector with reinforcement learning (Chen and Bansal, 2018), which has high training variance and low diversity on content selection.

In this paper, we treat the content selection as latent variables and train with amortized variational inference (Kingma and Welling, 2014; Mnih and Gregor, 2014). This provides a lower training variance than Reinforce-select. The selector and generator are co-trained within the same objective, the generations are thus more faithful to the selected contents than Bottom-up methods. Our model is task-agnostic, end-to-end trainable and can be seamlessly inserted into any encoder-decoder architecture. On both the data-to-text and headline generation task, we show our model outperforms others regarding content-level diversity and controllability while maintaining comparable performance. The performance/controllability tradeoff can be effectively adjusted by adjusting a single hyperparameter in the training stage, which constrains an upper bound of the conditional mutual information (CMI) between the selector and generated text (Alemi et al., 2018; Zhao et al., 2018). A higher CMI leads to stronger controllability with a bit more risk of text disfluency.

In summary, our contributions are **(1)** systematically studying the problem of controllable content selection for Enc-Dec text generation, **(2)** proposing a task-agnostic training framework achieving promising results and **(3)** introducing an effective way to achieve the trade-off between performance and controllability.

2 Background and Notation

Let X, Y denote a source-target pair. X is a sequence of $x_1, x_2, \dots, x_n$ and can be either some structured data or unstructured text/image depending on the task. Y corresponds to $y_1, y_2, \dots, y_m$ which is a text description of X. The goal of text generation is to learn a distribution $p(Y|X)$ to automatically generate proper text.

The Enc-Dec architecture handles this task with an encode-attend-decode process (Bahdanau et al., 2015; Xu et al., 2015). The encoder first encodes each x_i into a vector h_i. At each time step, the decoder pays attentions to some source embeddings and outputs the probability of the next token by $p(y_t|y_{1:t-1}, C_t)$. C_t is a weighted average of source embeddings:

$$
\begin{aligned}
C_t &= \sum_i \alpha_{t,i} h_i \\
\alpha_{t,i} &= \frac{e^{f(h_i,d_t)}}{\sum_j e^{f(h_j,d_t)}}
\end{aligned}
\tag{1}
$$

d_t is the hidden state of the decoder at time step t. f is a score function to compute the similarity between h_i and d_t (Luong et al., 2015).

3 Content Selection

Our goal is to decouple the content selection from the decoder by introducing an extra content selector. We hope the content-level diversity can be fully captured by the content selector for a more interpretable and controllable generation process. Following Gehrmann et al. (2018); Yu et al. (2018), we define content selection as a sequence labeling task. Let $\beta_1, \beta_2, \dots, \beta_n$ denote a sequence of binary selection masks. $\beta_i = 1$ if h_i is selected and 0 otherwise. β_i is assumed to be independent from each other and is sampled from a bernoulli distribution $\mathbf{B}(\gamma_i)$[2]. γ_i is the bernoulli parameter, which we estimate using a two-layer feedforward network on top of the source encoder. Text are generated by first sampling β from $\mathbf{B}(\gamma)$ to decide which content to cover, then decode with the conditional distribution $p_\theta(Y|X, \beta)$. The text is expected to faithfully convey all selected contents and drop unselected ones. Fig. 1 depicts this generation process. Note that the selection is based on the token-level *context-aware* embeddings h and will maintain information from the surrounding contexts. It encourages the decoder to stay faithful to the original information instead of simply fabricating random sentences by connecting the selected tokens. For each source-target pair, the ground-truth selection mask is unknown, so training is challenging. In the following session, we discuss several training possibilities and introduce the proposed model in detail.

[2] Devlin et al. (2019) have shown that excellent performance can be obtained by assuming such conditionally independence given a sufficiently expressive representation of x, though modelling a richer inter-label dependency is for sure beneficial (Lei et al., 2016; Nallapati et al., 2017).

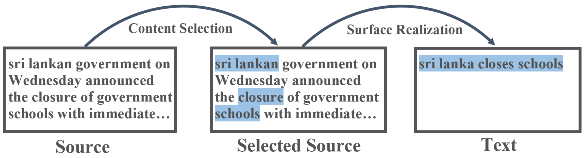

Figure 1: Model will select contents based on $\mathbf{B}(\gamma)$, then decode with $p_\theta(Y|X,\beta)$. Source-text pairs are available for training, but the ground-truth content selection for each pair is unknown.

3.1 Bottom-up

The most intuitive way is training the content selector to target some heuristically extracted contents. For example, we can train the selector to select overlapped words between the source and target (Gehrmann et al., 2018), sentences with higher tf-idf scores (Li et al., 2018) or identified image objects that appear in the caption (Wang et al., 2017). A standard encoder-decoder model is independently trained. In the testing stage, the prediction of the content selector is used to hard-mask the attention vector to guide the text generation in a bottom-up way. Though easy to train, Bottom-up generation has the following two problems: **(1)** The heuristically extracted contents might be coarse and cannot reflect the variety of human languages and **(2)** The selector and decoder are independently trained towards different objectives thus might not adapt to each other well.

β as Latent Variable: Another way is to treat β as a latent variable and co-train selector and generator by maximizing the marginal data likelihood. By doing so, the selector has the potential to automatically explore optimal selecting strategies best fit for the corresponding generator component.

With this in mind. We design $p_\theta(Y|X,\beta)$ by changing the original decoder in the following way: (1) We initialize hidden states of the decoder from a mean pooling over selected contents to inform the decoder which contents to cover and (2) Unselected contents will be prohibited from being attended to:

$$\begin{aligned} d_0 &= \text{MLP}\left(\frac{1}{n}\left(\sum_i^n \beta_i h_i\right)\right) \\ \alpha_{t,i} &= \frac{e^{f(h_i,d_t)}\beta_i}{\sum_j e^{f(h_j,d_t)}\beta_j} \end{aligned} \tag{2}$$

d_0 is the initial decoder hidden state and MLP denotes multi-layer-perceptron.

Since computing the exact marginal likelihood $\log \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)} p_\theta(Y|X,\beta)$ requires enumerating over all possible combinations of β (complexity $\mathbf{O}(2^n)$), we need some way to efficiently estimate the likelihood.

3.2 Soft-Select

Soft-select falls back on a *deterministic* network to output the likelihood function's first-order Taylor series approximation expanded at $\mathbb{E}_{\beta\sim\mathbf{B}(\gamma)}\beta$:

$$\begin{aligned} &\log \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)} p_\theta(Y|X,\beta) \\ \approx &\log[p_\theta(Y|X,\gamma) + \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)}(\beta-\gamma)p'_\theta(Y|X,\gamma)] \\ = &\log p_\theta(Y|X,\gamma) \end{aligned}$$

By moving the expectation into the decoding function, we can deterministically compute the likelihood by setting $\beta_i = \gamma_i$, reducing complexity to $\mathbf{O}(1)$. Each attention weight will first be "soft-masked" by γ before being passed to the decoder. soft-select is fully differentiable and can be easily trained by gradient descent. However, this soft-approximation is normally inaccurate, especially when $\mathbf{B}(\gamma)$ has a high entropy, which is common in one-to-many text generation tasks. The gap between $\log \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)} p_\theta(Y|X,\beta)$ and $\log p_\theta(Y|X, \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)})$ will be large (Ma et al., 2017; Deng et al., 2018). In practice, this would lead to unrealistic generations when sampling β from the deterministically trained distribution.

3.3 Reinforce-Select

Reinforce-select (RS) (Ling and Rush, 2017; Chen and Bansal, 2018) utilizes reinforcement learning to approximate the marginal likelihood. Specifically, it is trained to maximize a lower bound of the likelihood by applying the Jensen inequalily:

$$\log \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)} p_\theta(Y|X,\beta) \geq \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)} \log p_\theta(Y|X,\beta)$$

The gradient to γ is approximated with Monte-Carlo sampling by applying the REINFORCE algorithm (Williams, 1992; Glynn, 1990). To speed up convergence, we pre-train the selector by some distant supervision, which is a common practice in reinforcement learning. REINFORCE is unbiased but has a high variance. Many research have proposed sophisticated techniques for variance reduction (Mnih and Gregor, 2014; Tucker et al., 2017; Grathwohl et al., 2018). In text generation, the high-variance problem is aggravated because there exists multiple valid selections. Accurately estimating the likelihood becomes difficult. Another issue is its tendency to avoid stochasticity (Raiko et al., 2015), which we will show in Sec 5.2 that it results in low content-level diversity.

3.4 Variational Reinforce-Select

We propose Variational Reinforce-Select (VRS) which applies variational inference (Kingma and Welling, 2014) for variance reduction. Instead of directly integrating over $\mathbf{B}(\gamma)$, it imposes a proposal distribution q_ϕ for importance sampling. The marginal likelihood is lower bounded by:

$$\begin{aligned}
&\log \mathbb{E}_{\beta\sim\mathbf{B}(\gamma)} p_\theta(Y|X,\beta)\\
&= \log \mathbb{E}_{\beta\sim q_\phi} \frac{p_\theta(Y,\beta|X)}{q_\phi(\beta)}\\
&\geq \mathbb{E}_{\beta\sim q_\phi} \log \frac{p_\theta(Y,\beta|X)}{q_\phi(\beta)}\\
&= \mathbb{E}_{\beta\sim q_\phi} \log p_\theta(Y|X,\beta) - KL(q_\phi||\mathbf{B}(\gamma))
\end{aligned} \tag{3}$$

By choosing a proper q_ϕ, the bound will be improved and the variance can be largely reduced compared with REINFORCE. If q_ϕ equals the posterior distribution $p_\theta(\beta|X,Y)$, the bound is tight and the variance would be zero (Mnih and Rezende, 2016). We define $q_\phi(\beta|X,Y)$ as a mean-field distribution parameterized by a set of global parameters ϕ to approach the true posterior distribution. ϕ, θ and γ are simultaneously trained by minimizing the last tine of Eq. 3. $q_\phi(\beta|X,Y)$ also allows us to further perform posterior inference: Given an arbitrary text Y for a source X, we can infer which source contents are included in Y (An example is given in Appendix C).

In Eq.3, the KL divergence term can be computed analytically. As for the independence assumption, it can be summed over each individual β_i. The likelihood term is differentiable to θ but not to ϕ, we estimate the gradient to ϕ in Eq 3 by applying the REINFORCE estimator:

$$\begin{aligned}
&\nabla_\phi \mathbb{E}_{\beta\sim q_\phi} \log p_\theta(Y|X,\beta) =\\
&\mathbb{E}_{\beta\sim q_\phi} \nabla_\phi \log q_\phi(\beta|X,Y)(\log p_\theta(Y|X,\beta) - B)
\end{aligned}$$

B is the control variate (Williams, 1992). The optimal B would be (Weaver and Tao, 2001):

$$B^* = \mathbb{E}_{\beta\sim q_\phi} \log p_\theta(Y|X,\beta)$$

which we set as a soft-select approximation:

$$B = \log p_\theta(Y|X, \mathbb{E}_{\beta\sim q_\phi}\beta)$$

We estimate Eq. 3.4 with a single sample from q_ϕ for efficiency. Though multiple-sample could potentially further tighten the bound and reduce the variance (Burda et al., 2016; Lawson et al., 2018; Tucker et al., 2019), it brings significant computational overhead, especially in text generation tasks where the whole sentence needs to be decoded.

3.5 Degree of Controllability

In practice, when treating content selection as latent variables, the model tends to end up with a trivial solution of always selecting all source tokens (Shen et al., 2018a; Ke et al., 2018). This behavior is understandable since Eq. 2 strictly masks unselected tokens. Wrongly unselecting one token will largely deteriorate the likelihood. Under the maximum likelihood (MLE) objective, this high risk pushes the selector to take a conservative strategy of always keeping all tokens, then the whole model degenerates to the standard Enc-Dec and the selection mask loses effects on the generation. Usually people apply a penalty term to the selecting ratio when optimizing the likelihood:

$$\mathcal{L} + \lambda\big|(\bar{\gamma} - \alpha)\big| \tag{4}$$

$\mathcal{L}$ is the MLE loss function, $\bar{\gamma}$ is the mean of γ and α is the target selecting ratio. This forces the selector to select the most important α tokens for each source input instead of keeping all of them.

In our VRS model, we can easily adjust the degree of controllability by limiting an upper bound of the conditional mutual information (CMI) $I(\beta,Y|X)$ (Zhao et al., 2018). Specifically, we can change our objective into:

$$\begin{aligned}
&\max_{\phi,\theta,\gamma} \mathbb{E}_{\beta\sim q_\phi} \log p_\theta(Y|X,\beta)\\
&\quad -\lambda\big|KL(q_\phi||\mathbf{B}(\gamma)) - \epsilon)\big|
\end{aligned} \tag{5}$$

λ is a fixed lagrangian multiplier. Eq. 5 can be proved equal to maximum likelihood with the constraint $I(\beta,Y|X) = \epsilon$ given proper λ (Alemi et al., 2018). A higher ϵ indicates β has more influences to Y (higher controllability) while always safely selecting all tokens will lead $I(\beta,Y|X) = 0$.[3] It is preferred over Eq. 4 because (a) CMI directly considers the dependency between the selection and *multiple*-possible text while limiting the ratio aims at finding the *single* most salient parts for each source. (b) Unlike CMI, limiting the ratio is coarse. It considers only the total selected size and ignores its internal distribution.

[3]We also tried adding a coverage constraint to ensure the decoder covers all the selected tokens (Wen et al., 2015; Wang et al., 2019), but we find it brings no tangible help since a higher CMI can already discourage including redundant tokens into the selection.

Algorithm 1 Variational Reinforce-Select (VRS)

Parameters: θ, ϕ, γ
$pretrain \leftarrow$ TRUE
repeat
  Sample X,Y from the corpus;
  Encode X into $(h_1, h_2, \ldots, h_n)$;
  if $pretrain$ **then**
    Update ϕ with distant supervision;
    Update θ, γ by $\nabla_{\theta,\gamma}$Eq. 3;
  else
    Update θ, γ, ϕ by $\nabla_{\theta,\gamma,\phi}$Eq. 5;
  end if
  $pretrain \leftarrow$ FALSE if Eq. 3 degrades
until convergence and $pretrain$ is False

In practice, we can set ϵ to adjust the degree of controllability we want. Later we will show it leads to a trade-off with performance. The final algorithm is detailed in Algorithm 1. To keep fairness, we trian RS and VRS with the same control variate and pre-training strategy.[4]

4 Related Work

Most content selection models train the selector with heuristic rules (Hsu et al., 2018; Li et al., 2018; Yu et al., 2018; Gehrmann et al., 2018; Yao et al., 2019; Moryossef et al., 2019), which fail to fully capture the relation between selection and generation. Mei et al. (2016); Zhou et al. (2017); Lin et al. (2018); Li et al. (2018) "soft-select" word or sentence embeddings based on a gating function. The output score from the gate is a deterministic vector without any probabilistic variations, so controlling the selection to generate diverse text is impossible. Very few works explicitly define a bernoulli distribution for the selector, then train with the REINFORCE algorithm (Ling and Rush, 2017; Chen and Bansal, 2018), but the selection targets at a high recall regardless of the low precision, so the controllability over generated text is weak. Fan et al. (2018) control the generation by manually concatenating entity embeddings, while our model is much more flexible by explicitly defining the selection probability over all source tokens.

Our work is closely related with learning discrete representations with variational inference (Wen et al., 2017; van den Oord et al., 2017; Kaiser et al., 2018; Lawson et al., 2018), where we treat content selection as the latent representation. Limiting the KL-term is a common technique to deal with the "posterior collapse" problem (Kingma et al., 2016; Yang et al., 2017; Shen et al., 2018b). We adopt a similar approach and use it to further control the selecting strategy.

5 Experiments

For the experiments, we focus on comparing (1) Bottom-up generation (Bo.Up.), (2) soft-select (SS), (3) Reinforce-select (RS) and (4) Variational-Reinforce-select (VRS) regarding their performance on content selection. SS and RS are trained with the selecting ratio constraint in Eq. 4. For the SS model, we further add a regularization term to encourage the maximum value of γ to be close to 1 as in Mei et al. (2016). We first briefly introduce the tasks and important setup, then present the evaluation results.

5.1 Tasks and Setup

We test content-selection models on the headline and data-to-text generation task. Both tasks share the same framework with the only difference of source-side encoders.

Headline Generation: We use English Gigaword preprocessed by Rush et al. (2015), which pairs first sentences of news articles with their headlines. We keep most settings same as in Zhou et al. (2017), but use a vocabulary built by byte-pair-encoding (Sennrich et al., 2016). We find it speeds up training with superior performance.

Data-to-Text Generation: We use the Wikibio dataset (Lebret et al., 2016). The source is a Wikipedia infobox and the target is a one-sentence biography description. Most settings are the same as in Liu et al. (2018), but we use a bi-LSTM encoder for better performance.

Heuristically extracted content: This is used to train the selector for bottom up models and pre-train the RS and VRS model. For wikibio, we simply extract overlapped words between the source and target. In Gigaword, as the headline is more abstractive, we select the closest source word for each target word in the embedding space. Stop words and punctuations are prohibited from being selected.

Choice of α/ϵ: As seen in Sec 3.5, we need to set the hyperparameter α for RS/SS and ϵ for

[4]The only extra parameter of VRS is ϕ which is a simple MLP structure. The actual training complexity is similar to RS because they both use the REINFORCE algorithm for gradient estimation.

Gigaword	Oracle upper bound of ROUGE-1	ROUGE-2	ROUGE-L	% unique Generation	% unique Mask	% Effect of Selector	Entropy of Selector
Bo.Up.	42.61	22.32	38.37	84.28	95.87	87.91	0.360
SS	33.15	14.63	30.68	82.06	**96.23**	85.27	**0.392**
RS	36.62	18.34	34.60	3.01	6.23	48.31	0.018
VRS	**54.73**	**33.28**	**51.62**	**89.23**	92.51	**96.45**	0.288
Wikibio	ROUGE-4	BLEU-4	NIST	Generation	Mask	Selector	Selector
Bo.Up.	47.28	49.95	11.06	31.57	77.42	40.78	0.177
SS	41.73	43.94	9.82	**63.09**	**89.42**	70.55	**0.355**
RS	44.07	46.89	10.31	4.55	43.83	10.38	0.105
VRS	**52.41**	**55.03**	**11.89**	57.62	77.83	**74.03**	0.181

Table 2: Diversity of content selection. The % effect of selector is defined as the ratio of unique generation and mask, which reflects the rate that changing the selector will lead to corresponding changes of the generated text.

VRS. α corresponds to the selecting ratio. We set them as $\alpha = 0.35$ for Wikibio and 0.25 for Gigaword. The value is decided by running a human evaluation to get the empirical estimation. To keep comparison fairness, we tune ϵ to make VRS select similar amount of tokens with RS. The values we get are $\epsilon = 0.15n$ for Wikibio and $\epsilon = 0.25n$ for Gigaword. n is the number of source tokens. [5]

5.2 Results and Analysis

Ideally we would expect the learned content selector to (1) have reasonable diversity so that text with various contents can be easily sampled, (2) properly control the contents described in the generated text and (3) not hurt performance. The following section will evaluate these three points in order.

Diversity: We first look into the diversity of content selection learned by different models. For each test data, 50 selection masks are randomly sampled from the model's learned distribution. Greedy decoding is run to generate the text for each mask. We measure the entropy of the selector, proportion of unique selection masks and generated text in the 50 samples. We further define the "effect" of the selector as the ratio of sampled unique text and mask. This indicates how often changing the selection mask will also lead to a change in the generated text. The results are averaged over all test data. Following Rush et al. (2015) and Lebret et al. (2016), we measure the quality of generated text with ROUGE-1, 2, L F-score for Gigaword and ROUGE-4, BLEU-4, NIST for Wikibio. As there is only one reference text for each source, we report an oracle upper bound of these scores by assuming an "oracle" that can choose the best text among all the candidates (Mao et al., 2015; Wang et al., 2017). Namely, out of each 50 sampled text, we pick the one with the maximum metric score. The final metric score is evaluated on these "oracle" picked samples. The intuition is that if the content selector is properly trained, at least one out of the 50 samples should describe similar contents with the reference text, the metric score between it and the reference text should be high. Table 2 lists the results. We can have the following observations:

- RS model completely fails to capture the content-level diversity. Its selector is largely deterministic, with a lowest entropy value among all models. In contrast, the selector from SS, VRS and Bo.Up. have reasonable diversity, with over 90% and 75% unique selection masks for Gigaword and Wikibio respectively.
- The selector from VRS has the strongest effect to the generator, especially on the Gigaword data where modifying the content selection changes the corresponding text in more than 95% of the cases. RS has the lowest effect value, which indicates that even with the selecting ratio constraint, its generator still ignores the selection mask to a large extent.
- The oracle metric score of VRS is much higher than the other two. This is beneficial when people want to apply the model to generate a few candidate text then hand-pick the suitable one. VRS has more potential than the other three to contain the expected text. SS performs worst. The gap between the

[5] ϵ corresponds to the KL divergence of the selection mask, which scales linearly with the number of source tokens, so we set it proportinally w.r.t. n.

soft approximation and the real distribution, as mentioned before, indeed results in a large drop of performance.

In short, compared with others, the content selector of VRS is *(1) diverse, (2) has stronger effect on the text generation and (3) with a larger potential of producing an expected text.*

Controllability: We have shown the content selector of VRS is diverse and has strong effect on the text generation. This section aims at examining whether such effect is desirable, i.e., whether the selector is able to properly control the contents described in the text. We measure it based on the self-bleu metric and a human evaluation.

The self-bleu metric measures the controllability by evaluating the "intra-selection" similarity of generated text. Intuitively, by fixing the selection mask, multiple text sampled from the decoder are expected to describe the same contents and thereby should be highly similar to each other. The decoder should only model surface-level diversity without further modifying the selected contents. With this in mind, for each test data, we randomly sample a selection mask from the selector's distribution, then fix the mask and run the decoder to sample 10 different text. The self-BLEU-1 score (Zhu et al., 2018) on the sampled text is reported, which is the average BLEU score between each text pair. A higher self-BLEU score indicates the sampled text are more similar with each other. The results are shown in Table 3. We can see generations from VRS have a clearly higher intra-selection similarity. SS performs even worse than RS, despite having a high effect score in Table 2. The selector from SS affects the generation in an undesirable way, which also explain why SS has a lowest oracle metric score though with a high score on content diversity and effect.

Method	Bo.Up.	SS	RS	VRS
Gigaword	46.58	37.20	48.13	**61.14**
Wikibio	38.30	13.92	25.99	**43.81**

Table 3: Self-Bleu score by fixing selection mask. Higher means better controllability of content selection

We further run a human evaluation to measure the text-content consistency among different models. 100 source text are randomly sampled from the human-written DUC 2004 data for task 1&2 (Over et al., 2007). Bo.Up, SS, RS and VRS are applied to generate the target text by first sampling a selection mask, then run beam search decoding with beam size 10. We are interested in seeing (1) if multiple generations from the same selection mask are paraphrases to each other (intra-consistent) and (2) if generations from different selection masks do differ in the content they described (inter-diverse). The results in Table 4 show that VRS significantly outperforms the other two in both intra-consistency and inter-diversity. RS has the lowest score on both because the selector has very weak effects on the generation as measured in the last section. Bo.Up and SS lay between them. Overall VRS is able to *maintain the highest content-text consistency* among them.

Method	Fluency	intra-consistency	inter-diversity
Reference	0.96	-	-
Enc-Dec	0.83	-	-
Bo.Up.	0.46	0.48	0.61
SS	0.27	0.41	0.54
RS	**0.78**	0.39	0.47
VRS	0.74	**0.72**	**0.87**

Table 4: Human evaluation on fluency, intra-consistency and inter-diversity of content selection on DUC 2004.

Method	R-1	R-2	R-L	%Word
Zhou et al. (2017)	36.15	17.54	33.63	100
Enc-Dec	35.92	17.43	33.42	100
SS	20.35	4.78	16.53	24.82
Bo.Up	28.17	10.32	26.68	24.54
RS	35.45	16.38	32.71	25.12
VRS($\epsilon = 0$)-pri	**36.42**	**17.81**	**33.86**	78.63
VRS($\epsilon = 0.25$)-pri	34.26	15.11	31.69	24.36
VRS($\epsilon = 0$)-post	37.14	18.03	34.26	78.66
VRS($\epsilon = 0.25$)-post	**56.72**	**33.24**	**51.88**	24.53

Table 5: Gigaword best-select results. Larger ϵ leads to more controllable selector with a bit degrade of performance. (-post) means selecting from the posterior $q_\phi(\beta|X, Y)$, (-pri) is from the prior $\mathbf{B}(\gamma_i)$.

Performance & **Trade-off**: To see if the selector affects performance, we also ask human annotators to judge the text fluency. The fluency score is computed as the average number of text being judged as fluent. We include generations from the standard Enc-Dec model. Table 4 shows the best fluency is achieved for Enc-Dec. Imposing

Method	R-4	B-4	NIST	%Word
Liu et al. (2018)	41.65	44.71		100
Enc-Dec	42.07	44.80	9.82	100
SS	5.10	5.73	0.24	35.12
Bo.Up	8.07	9.52	0.42	38.79
RS	42.64	45.08	10.01	34.53
VRS($\epsilon = 0$)-pri	**43.01**	**46.01**	**10.24**	84.56
VRS($\epsilon = 0.15$)-pri	42.13	44.51	9.84	34.04
VRS($\epsilon = 0$)-post	43.84	46.60	10.27	85.34
VRS($\epsilon = 0.15$)-post	**49.68**	**52.26**	**11.48**	34.57

Table 6: Wikibio best-select results.

a content selector always affects the fluency a bit. The main reason is that when the controllability is strong, the change of selection will directly affect the text realization so that a tiny error of content selection might lead to unrealistic text. If the selector is not perfectly trained, the fluency will inevitably be influenced. When the controllability is weaker, like in RS, the fluency is more stable because it will not be affected much by the selection mask. For SS and Bo.Up, the drop of fluency is significant because of the gap of soft approximation and the independent training procedure. In general, VRS does properly decouple content selection from the enc-dec architecture, with only tiny degrade on the fluency.

Table 5/6 further measure the metric scores on Gigaword/Wikibio by decoding text from the best selection mask based on the selector's distribution (set $\beta_i = 1$ if $\mathbf{B}(\gamma_i) > 0.5$ and 0 otherwise). We include results from VRS model with $\epsilon = 0$, which puts no constraint on the mutual information. We further report the score by generating the best selection mask from the learned posterior distribution $q_\phi(\beta|X, Y)$ for VRS model. Two current SOTA results from Zhou et al. (2017) and Liu et al. (2018) and the proportion of selected source words for each model are also included. We have the following observations:

- As the value of ϵ decreases, the performance of VRS improves, but the selector loses more controllability because the model tends to over-select contents (over 75% source words selected). The text-content consistency will become low.

- Increasing ϵ sacrifices a bit performance, but still comparable with SOTA. Especially on Wikibio where the performance drop is minor. The reason should be that Wikibio is relatively easier to predict the selection but Gigaword has more uncertainty.

- Increasing ϵ improves the accuracy of the posterior selection. This would be useful when we want to perform posterior inference for some source-target pair.

- Setting $\epsilon = 0$ can actually outperform SOTA seq2seq which keeps all tokens, suggesting it is still beneficial to use the VRS model even if we do not care about the controllability.

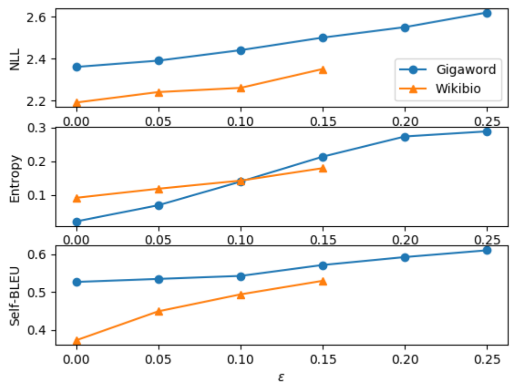

Figure 2: Negative log likelihood (NLL), selection entropy and self-BLEU as ϵ changes. NLL and self-bleu on Wikibio are added by 1 for better visualization. Lower NLL suggests higher performance. Higher entropy/self-BLEU means higher diversity/controllability.

Figure 2 visualizes how changing the value of ϵ affects the negative log likelihood (NLL), entropy of the selector and self-bleu score, which roughly correlates with performance, diversity and controllability. NLL is evaluated based on the lower bound in Eq 3 (Sohn et al., 2015). We can see as ϵ increases, the performance decreases gradually but the content selection gains more diversity and controllability. In practice we can tune the ϵ value to achieve a trade-off.

Generation Example: Figure 3 shows some examples from Gigaword. As can be seen, decodings from the VRS model are largely consistent with each other, in most cases only replacing one or two words with corresponding synonyms. Samples are able to faithfully convey all selected contents. In contrast, generations from SS. Bo.Up. and RS are unpreditable, differing in both selected

Source: indian prime minister p.v. narasimha rao 's promise of more autonomy for troubled kashmir and his plea for early state elections has sparked a violent reaction from provincial moslem and opposition parties .
Samples from SS:
t1: indian indian calls for end to violence in kashmir .
t2: indian pm calls for end to violence in afghanistan .
t3: indian pm calls for boycott of pakistan 's ruling party .
Samples from Bo.Up:
t1: india promises more autonomous more autonomy .
t2: indian pm promises autonomy for kashmir autonomy .
t3: indian pm 's promise sparks violent reaction .
Samples from RS:
t1: indian pm 's kashmir promises sparks violent reaction.
t2: indian pm 's promise sparks violent reaction .
t3: kashmir parties blast pm 's promise .
Samples from VRS:
t1: indian pm 's promise on kashmir sparks uproar .
t2: indian pm 's promise on kashmir sparks protests .
t3: indian pm 's promise for kashmir sparks controversy .

Source: factory orders for manufactured goods rose #.# percent in september , the commerce department said here thursday .
Samples from SS:
t1: u.s. consumer confidence down in january in january.
t2: u.s. wholesale prices up #.# percent in october .
t3: u.s. jobless rate rises to #.# percent in march .
Samples from Bo.Up.:
t1: september u.s. factory orders up #.# percent .
t2: september u.s. factory orders increase .
t3: factory orders up in september .
Samples from RS:
t1: u.s. factory orders up #.# percent in september .
t2: factory orders for manufactured goods rise .
t3: factory orders up in september from the year .
Samples from VRS:
t1: september factory orders up #.# percent .
t2: september factory orders rise #.# percent .
t3: september factory orders increase #.# pct .

Figure 3: Text generation examples from Gigaword. Highlighted words are selected. t1-3 are sampled from the decoder based on the selected content. Generations from VRS are more faithful to selected contents.

contents and also the way of saying. SS and Bo.Up also suffer more from the text disfluency. The generations from them are largely uncertain.

6 Conclusion

In this paper, we tackle the unaddressed problem of controllable content selection in text generation. We propose a general framework based on variational inference that can be potentiall applied to arbitrary tasks. On both the headline generation and data-to-text tasks, our model outperforms state-of-the-art models regarding the diversity and controllability of content selection. We further introduce an effective way to achieve a performance/controllability trade-off, which can be easily tuned to meet specific requirement.

Acknowledgments

We thank anonymous reviewers for valuable comments, thank Aditya Mogadala, Shun Kiyono, Thomas Mclachlan and other members of the LIAT team at RIKEN AIP for useful discussions. Xiaoyu Shen is supported by IMPRS-CS fellowship. The work of J. Suzuki was partly supported by JSPS KAKENHI Grant Number JP19104418 and AIRPF Grant Number 30AI036-8. This work is also partially funded by DFG collaborative research center SFB 1102.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. 2018. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 159–168, Stockholmsmssan, Stockholm Sweden. PMLR.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. 2016. Importance weighted autoencoders. In *International Conference on Learning Representations*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686. Association for Computational Linguistics.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2018. Show, control and tell: A framework for generating controllable and grounded captions. *arXiv preprint arXiv:1811.10652*.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander M Rush. 2018. Latent alignment and variational attention. *Nips*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *EMNLP*.

Peter W. Glynn. 1990. Likelihood ratio gradient estimation for stochastic systems. *Commun. ACM*, 33(10):75–84.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud. 2018. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*.

Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12):e28210.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141. Association for Computational Linguistics.

Peter Jackson. 1998. *Introduction to expert systems*. Addison-Wesley Longman Publishing Co., Inc.

Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2390–2399, Stockholmsmssan, Stockholm Sweden. PMLR.

Nan Rosemary Ke, Konrad Żołna, Alessandro Sordoni, Zhouhan Lin, Adam Trischler, Yoshua Bengio, Joelle Pineau, Laurent Charlin, and Christopher Pal. 2018. Focused hierarchical RNNs for conditional sequence processing. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2554–2563, Stockholmsmssan, Stockholm Sweden. PMLR.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Dieterich Lawson, Chung-Cheng Chiu, George Tucker, Colin Raffel, Kevin Swersky, and Navdeep Jaitly. 2018. Learning hard alignments with variational inference. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5799–5803. IEEE.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with explicit information selection modeling. *EMNLP*.

Junyang Lin, Xu SUN, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 163–169. Association for Computational Linguistics.

Jeffrey Ling and Alexander Rush. 2017. Coarse-to-fine attention models for document summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 33–42.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4881–4888. AAAI Press.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Xuezhe Ma, Yingkai Gao, Zhiting Hu, Yaoliang Yu, Yuntian Deng, and Eduard Hovy. 2017. Dropout with expectation-linear regularization. In *International Conference on Learning Representations*.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*.

Hongyuan Mei, TTI UChicago, Mohit Bansal, and Matthew R Walter. 2016. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *Proceedings of NAACL-HLT*, pages 720–730.

Andriy Mnih and Karol Gregor. 2014. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1791–II–1799. JMLR.org.

Andriy Mnih and Danilo J. Rezende. 2016. Variational inference for monte carlo objectives. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2188–2196. JMLR.org.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. *NAACL*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Preksha Nema, Mitesh M Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1063–1072.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6306–6315. Curran Associates, Inc.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. *NAACL*.

Guanghui Qin, Jin-Ge Yao, Xuening Wang, Jinpeng Wang, and Chin-Yew Lin. 2018. Learning latent semantic annotations for grounding natural language to structured data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3761–3771.

Tapani Raiko, Mathias Berglund, Guillaume Alain, and Laurent Dinh. 2015. Techniques for learning binary stochastic feedforward neural networks. *ICLR*.

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

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018a. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4345–4352. International Joint Conferences on Artificial Intelligence Organization.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018b. Improving variational encoder-decoders in dialogue generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 79. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. 2019. Doubly reparameterized gradient estimators for monte carlo objectives. In *International Conference on Learning Representations*.

George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. 2017. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2627–2636. Curran Associates, Inc.

Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766.

Wentao Wang, Zhiting Hu, Zichao Yang, Haoran Shi, Frank Xu, and Eric Xing. 2019. Toward unsupervised text content manipulation. *arXiv preprint arXiv:1901.09501*.

Lex Weaver and Nigel Tao. 2001. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 538–545, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3732–3741, International Convention Centre, Sydney, Australia. PMLR.

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

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2018. Learning neural templates for text generation. *EMNLP*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. *ICML*.

Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. *AAAI*.

Naitong Yu, Jie Zhang, Minlie Huang, and Xiaoyan Zhu. 2018. An operation network for abstractive sentence compression. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1065–1076. Association for Computational Linguistics.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. 2018. The information autoencoding family: A lagrangian perspective on latent variable generative models. *UAI*.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104. 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*, pages 1097–1100. ACM.