Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs

Dong Bok Lee[1*] **Seanie Lee**[1,3*] **Woo Tae Jeong**[3] **Donghwan Kim**[3] **Sung Ju Hwang**[1,2]
KAIST[1], AITRICS[2], 42Maru Inc.[3], South Korea
{markhi,lsnfamily02,sjhwang82}@kaist.ac.kr
{wtjeong,scissors}@42maru.com

Abstract

One of the most crucial challenges in question answering (QA) is the scarcity of labeled data, since it is costly to obtain question-answer (QA) pairs for a target text domain with human annotation. An alternative approach to tackle the problem is to use automatically generated QA pairs from either the problem context or from large amount of unstructured texts (e.g. Wikipedia). In this work, we propose a hierarchical conditional variational autoencoder (HCVAE) for generating QA pairs given unstructured texts as contexts, while maximizing the mutual information between generated QA pairs to ensure their consistency. We validate our **Info**rmation Maximizing **H**ierarchical **C**onditional **V**ariational **A**uto**E**ncoder (**Info-HCVAE**) on several benchmark datasets by evaluating the performance of the QA model (BERT-base) using only the generated QA pairs (QA-based evaluation) or by using both the generated and human-labeled pairs (semi-supervised learning) for training, against state-of-the-art baseline models. The results show that our model obtains impressive performance gains over all baselines on both tasks, using only a fraction of data for training. [1]

1 Introduction

Extractive Question Answering (QA) is one of the most fundamental and important tasks for natural language understanding. Thanks to the increased complexity of deep neural networks and use of knowledge transfer from the language models pretrained on large-scale corpora (Peters et al., 2018; Devlin et al., 2019; Dong et al., 2019), the state-of-the-art QA models have achieved human-level performance on several benchmark datasets (Rajpurkar et al., 2016, 2018). However, what is also crucial to the success of the recent data-driven models, is the availability of large-scale QA datasets. To deploy the state-of-the-art QA models to real-world applications, we need to construct high-quality datasets with large volumes of QA pairs to train them; however, this will be costly, requiring a massive amount of human efforts and time.

Paragraph (Input)	Philadelphia has more murals than any other u.s. city, thanks in part to the 1984 creation of the department of recreation's mural arts program, . . . The program has funded more than 2,800 murals
Q1	which city has more murals than any other city?
A1	philadelphia
Q2	why philadelphia has more murals?
A2	the 1984 creation of the department of recreation's mural arts program
Q3	when did the department of recreation' s mural arts program start ?
A3	1984
Q4	how many murals funded the graffiti arts program by the department of recreation?
A4	more than 2,800

Table 1: An example of QA pairs generated with our framework. The paragraph is an extract from Wikipedia provided by Du and Cardie (2018). For more examples, please see Appendix D.

Question generation (QG), or *Question-Answer pair generation (QAG)*, is a popular approach to overcome this data scarcity challenge. Some of the recent works resort to semi-supervised learning, by leveraging large amount of unlabeled text (e.g. Wikipedia) to generate synthetic QA pairs with the help of QG systems (Tang et al., 2017; Yang et al., 2017; Tang et al., 2018; Sachan and Xing, 2018). However, existing QG systems have overlooked an important point that generating QA pairs from a context consisting of unstructured texts, is essentially a *one-to-many* problem. Sequence-to-sequence models are known to generate generic sequences (Zhao et al., 2017a) without much variety, as they are trained with maximum likelihood estimation. This is highly suboptimal for QAG

* Equal contribution

[1] The generated QA pairs and the code can be found at https://github.com/seanie12/Info-HCVAE

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 208–224
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

since the contexts given to the model often contain richer information that could be exploited to generate multiple QA pairs.

To tackle the above issue, we propose a novel probabilistic deep generative model for QA pair generation. Specifically, our model is a hierarchical conditional variational autoencoder (HCVAE) with two separate latent spaces for question and answer conditioned on the context, where the answer latent space is additionally conditioned on the question latent space. During generation, this hierarchical conditional VAE first generates an answer given a context, and then generates a question given both the answer and the context, by sampling from both latent spaces. This probabilistic approach allows the model to generate diverse QA pairs focusing on different parts of a context at each time.

Another crucial challenge of the QG task is to ensure the *consistency* between a question and its corresponding answer, since they should be semantically dependent on each other such that the question is answerable from the given answer and the context. In this paper, we tackle this consistency issue by maximizing the mutual information (Belghazi et al., 2018; Hjelm et al., 2019; Yeh and Chen, 2019) between the generated QA pairs. We empirically validate that the proposed mutual information maximization significantly improves the QA-pair consistency. Combining both the hierarchical CVAE and the InfoMax regularizer together, we propose a novel probabilistic generative QAG model which we refer to as **Info**rmation Maximizing **H**ierarchical **C**onditional **V**ariational **A**uto**E**ncoder (**Info-HCVAE**). Our Info-HCVAE generates diverse and consistent QA pairs even from a very short context (see Table 1).

But how should we quantitatively measure the quality of the generated QA pairs? Popular evaluation metrics (e.g. BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2002), METEOR (Banerjee and Lavie, 2005)) for text generation only tell how similar the generated QA pairs are to Ground-Truth (GT) QA pairs, and are not directly correlated with their actual quality (Nema and Khapra, 2018; Zhang and Bansal, 2019). Therefore, we use the **QA**-based **E**valuation (**QAE**) metric proposed by Zhang and Bansal (2019), which measures how well the generated QA pairs match the distribution of GT QA pairs. Yet, in a semi-supervised learning setting where we already have GT labels, we need novel QA pairs that are different from GT QA pairs for the additional QA pairs to be truly effective. Thus, we propose a novel metric, **R**everse **QAE** (**R-QAE**), which is low if the generated QA pairs are novel and diverse.

We experimentally validate our QAG model on SQuAD v1.1 (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019), and TriviaQA (Joshi et al., 2017) datasets, with both QAE and R-QAE using BERT-base (Devlin et al., 2019) as the QA model. Our QAG model obtains high QAE and low R-QAE, largely outperforming state-of-the-art baselines using a significantly smaller number of contexts. Further experimental results for semi-supervised QA on the three datasets using the SQuAD as the labeled dataset show that our model achieves significant improvements over the state-of-the-art baseline (+2.12 on SQuAD, +5.67 on NQ, and +1.18 on Trivia QA in EM).

Our contribution is threefold:

- We propose a novel hierarchical variational framework for generating diverse QA pairs from a single context, which is, to our knowledge, the first probabilistic generative model for question-answer pair generation (QAG).
- We propose an InfoMax regularizer which effectively enforces the consistency between the generated QA pairs, by maximizing their mutual information. This is a novel approach in resolving consistency between QA pairs for QAG.
- We evaluate our framework on several benchmark datasets by either training a new model entirely using generated QA pairs (QA-based evaluation), or use both ground-truth and generated QA pairs (semi-supervised QA). Our model achieves impressive performances on both tasks, largely outperforming existing QAG baselines.

2 Related Work

Question and Question-Answer Pair Generation Early works on Question Generation (QG) mostly resort to rule-based approaches (Heilman and Smith, 2010; Lindberg et al., 2013; Labutov et al., 2015). However, recently, encoder-decoder based neural architectures (Du et al., 2017; Zhou et al., 2017) have gained popularity as they outperform rule-based methods. Some of them use paragraph-level information (Du and Cardie, 2018; Song et al., 2018; Liu et al., 2019; Zhao et al., 2018; Kim et al., 2019; Sun et al., 2018) as additional information. Reinforcement learning is a popular approach to train the neural QG models, where the reward is defined as the evaluation metrics (Song et al., 2017; Kumar et al., 2018), or the QA accuracy/likelihood (Yuan et al., 2017; Hosking and Riedel, 2019; Zhang and Bansal, 2019). State-of-the-art QG models (Alberti et al., 2019; Dong et al., 2019; Chan and Fan, 2019) use pre-trained language models. Question-Answer Pair Generation (QAG) from contexts, which is our main target, is a relatively less explored topic tackled by only a few recent works (Du and Cardie, 2018; Alberti et al., 2019; Dong et al., 2019). To the best of our knowledge, we are the first to propose a probabilistic generative model for end-to-end QAG; Yao et al. (2018) use VAE for QG, but they do not tackle QAG. Moreover, we effectively resolve the QA-pair consistency issue by maximizing their mutual information with an InfoMax regularizer (Belghazi et al., 2018; Hjelm et al., 2019; Yeh and Chen, 2019), which is another contribution of our work.

Semi-supervised QA with QG With the help of QG models, it is possible to train the QA models in a semi-supervised learning manner to obtain improved performance. Tang et al. (2017) apply dual learning to jointly train QA and QG on unlabeled dataset. Yang et al. (2017) and Tang et al. (2018) train QG and QA in a GAN framework (Goodfellow et al., 2014). Sachan and Xing (2018) propose a curriculum learning to supervise the QG model to gradually generate difficult questions for the QA model. Dhingra et al. (2018) introduce a cloze-style QAG method to pretrain a QA model. Zhang and Bansal (2019) propose to filter out low-quality synthetic questions by the answer likelihood. While we focus on the answerable setting in this paper, few recent works tackle the unanswerable settings. Zhu et al. (2019) use neural networks to edit answerable questions into unanswerable ones, and perform semi-supervised QA. Alberti et al. (2019) and Dong et al. (2019) convert generated questions into unanswerable ones using heuristics, and filter or replace corresponding answers based on EM or F1.

Variational Autoencoders Variational autoencoders (VAEs) (Kingma and Welling, 2014) are probabilistic generative models used in a variety of natural language understanding tasks, including language modeling (Bowman et al., 2016), dialogue generation (Serban et al., 2017; Zhao et al., 2017b; Park et al., 2018; Du et al., 2018; Qiu et al., 2019), and machine translation (Zhang et al., 2016; Su et al., 2018; Deng et al., 2018). In this work, we propose a novel hierarchical conditional VAE framework with an InfoMax regularization for generating a pair of samples with high consistency.

3 Method

Our goal is to generate diverse and consistent QA pairs to tackle the data scarcity challenge in the extractive QA task. Formally, given a *context* $\mathbf{c}$ which contains M tokens, $\mathbf{c} = (c_1, \ldots, c_M)$, we want to generate QA pairs $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} = (x_1, \ldots, x_N)$ is the question containing N tokens and $\mathbf{y} = (y_1, \ldots, y_L)$ is its corresponding answer containing L tokens. We aim to tackle the QAG task by learning the conditional joint distribution of the question and answer given the context, $p(\mathbf{x}, \mathbf{y}|\mathbf{c})$, from which we can sample the QA pairs:

$$(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y}|\mathbf{c})$$

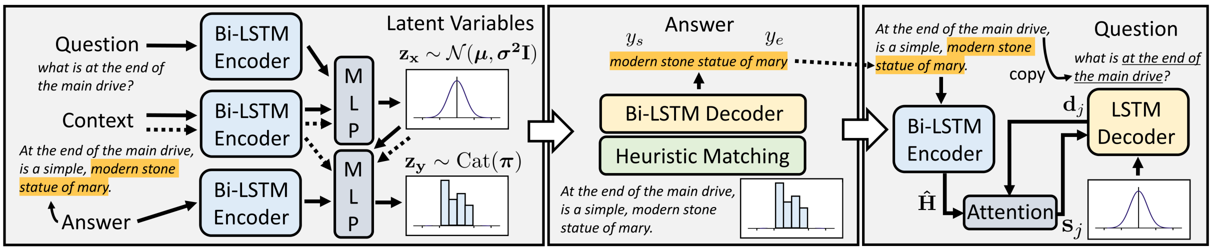
We estimate $p(\mathbf{x}, \mathbf{y}|\mathbf{c})$ with a probabilistic deep generative model, which we describe next.

3.1 Hierarchical Conditional VAE

We propose to approximate the unknown conditional joint distribution $p(\mathbf{x}, \mathbf{y}|\mathbf{c})$, with a variational autoencoder (VAE) framework (Kingma and Welling, 2014). However, instead of directly learning a common latent space for both question and answer, we model $p(\mathbf{x}, \mathbf{y}|\mathbf{c})$ in a hierarchical conditional VAE framework with a separate latent space for question and answer as follows:

$$\begin{aligned} &p_\theta(\mathbf{x}, \mathbf{y}|\mathbf{c}) \\ &= \int_{\mathbf{z_x}} \sum_{\mathbf{z_y}} p_\theta(\mathbf{x}|\mathbf{z_x}, \mathbf{y}, \mathbf{c}) p_\theta(\mathbf{y}|\, \mathbf{z_x}, \mathbf{z_y}, \mathbf{c}) \cdot \\ &\qquad\qquad p_\psi(\mathbf{z_y}|\mathbf{z_x}, \mathbf{c}) p_\psi(\mathbf{z_x}|\mathbf{c}) d\mathbf{z_x} \end{aligned}$$

where $\mathbf{z_x}$ and $\mathbf{z_y}$ are latent variables for question and answer respectively, and the $p_\psi(\mathbf{z_x}|\mathbf{c})$ and $p_\psi(\mathbf{z_y}|\mathbf{z_x}, \mathbf{c})$ are their conditional priors following an isotropic Gaussian distribution and a categorical distribution (Figure 1-(a)). We decompose the latent space of question and answer, since the answer is always a finite span of context $\mathbf{c}$, which can be modeled well by a categorical distribution, while a continuous latent space is a more appropriate choice for question since there could be unlimited valid questions from a single context. Moreover, we design the bi-directional dependency flow of joint distribution for QA. By leveraging hierarchical structure, we enforce the answer latent variables


Question
what is at the end of the main drive?
Context
At the end of the main drive, is a simple, modern stone statue of mary.
Answer
Bi-LSTM Encoder
Bi-LSTM Encoder
Bi-LSTM Encoder
MLP
MLP
Latent Variables
$\mathbf{z_x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$
$\mathbf{z_y} \sim \mathrm{Cat}(\boldsymbol{\pi})$
Answer
y_s
y_e
modern stone statue of mary
Bi-LSTM Decoder
Heuristic Matching
At the end of the main drive, is a simple, modern stone statue of mary.
At the end of the main drive, is a simple, modern stone statue of mary.
Question
copy
what is at the end of the main drive?
Bi-LSTM Encoder
$\mathbf{d}_j$
LSTM Decoder
$\hat{\mathbf{H}}$
Attention
$\mathbf{s}_j$

(a) Conditional Posterior / Prior Networks (b) Answer Generation Network (c) Question Generation Network

Figure 1: The conceptual illustration of the proposed HCVAE model encoding and decoding question and its corresponding answer jointly. The dashed line refers to the generative process of HCVAE.

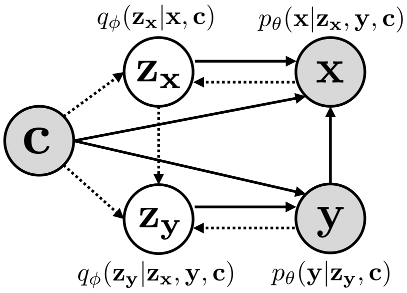

$q_\phi(\mathbf{z_x}|\mathbf{x},\mathbf{c})$
$p_\theta(\mathbf{x}|\mathbf{z_x},\mathbf{y},\mathbf{c})$
$\mathbf{z_x}$
$\mathbf{x}$
$\mathbf{c}$
$\mathbf{z_y}$
$\mathbf{y}$
$q_\phi(\mathbf{z_y}|\mathbf{z_x},\mathbf{y},\mathbf{c})$
$p_\theta(\mathbf{y}|\mathbf{z_y},\mathbf{c})$

Figure 2: The directed graphical model for HCVAE. The gray and white nodes denote observed and latent variables.

to be dependent on the question latent variables in $p_\psi(\mathbf{z_y}|\mathbf{z_x},\mathbf{c})$ and achieve the reverse dependency by sampling question $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z_x},\mathbf{y},\mathbf{c})$. We then use a variational posterior $q_\phi(\cdot)$ to maximize the Evidence Lower Bound (ELBO) as follows (The complete derivation is provided in **Appendix** A):

$$\begin{aligned}\log p_\theta(\mathbf{x},\mathbf{y}|\mathbf{c}) \geq\ & \mathbb{E}_{\mathbf{z_x}\sim q_\phi(\mathbf{z_x}|\mathbf{x},\mathbf{c})}[\log p_\theta(\mathbf{x}|\mathbf{z_x},\mathbf{y},\mathbf{c})] \\ & + \mathbb{E}_{\mathbf{z_y}\sim q_\phi(\mathbf{z_y}|\mathbf{z_x},\mathbf{y},\mathbf{c})}[\log p_\theta(\mathbf{y}|\mathbf{z_y},\mathbf{c})] \\ & - \mathrm{D_{KL}}[q_\phi(\mathbf{z_y}|\mathbf{z_x},\mathbf{y},\mathbf{c})||p_\psi(\mathbf{z_y}|\mathbf{z_x},\mathbf{c})] \\ & - \mathrm{D_{KL}}[q_\phi(\mathbf{z_x}|\mathbf{x},\mathbf{c})||p_\psi(\mathbf{z_x}|\mathbf{c})] \\ & =: \mathcal{L}_{\mathrm{HCVAE}}\end{aligned}$$

where θ, ϕ, and ψ are the parameters of the generation, posterior, and prior network, respectively. We refer to this model as a *Hierarchical Conditional Variational Autoencoder (HCVAE)* framework.

Figure 2 shows the directed graphical model of our HCVAE. The generative process is as follows:

1. Sample question L.V.: $\mathbf{z_x} \sim p_\psi(\mathbf{z_x} \mid \mathbf{c})$
2. Sample answer L.V.: $\mathbf{z_y} \sim p_\psi(\mathbf{z_y} \mid \mathbf{z_x},\mathbf{c})$
3. Generate an answer: $\mathbf{y} \sim p_\theta(\mathbf{y} \mid \mathbf{z_y},\mathbf{c})$
4. Generate a question: $\mathbf{x} \sim p_\theta(\mathbf{x} \mid \mathbf{z_x},\mathbf{y},\mathbf{c})$

Embedding We use the pre-trained word embedding network from BERT (Devlin et al., 2019) for posterior and prior networks, whereas the whole BERT is used as a contextualized word embedding model for the generative networks. For the answer encoding, we use a binary token type id of BERT. Specifically, we encode all context tokens as 0s, except for the tokens which are part of answer span (highlighted words of context in Figure 1-(a) or -(c)), which we encode as 1s. We then feed the sequence of the word token ids, token type ids, and position ids into the embedding layer to encode the answer-aware context. We fix all the embedding layers in HCVAE during training.

Prior Networks We use two different conditional prior networks $p_\psi(\mathbf{z_x}|\mathbf{c}), p_\psi(\mathbf{z_y}|\mathbf{z_x},\mathbf{c})$ to model context-dependent priors (the dashed lines in Figure 1-(a)). To obtain the parameters of isotropic Gaussian $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma}^2\mathbf{I})$ for $p_\psi(\mathbf{z_x}|\mathbf{c})$, we use a bi-directional LSTM (Bi-LSTM) to encode the word embeddings of the context into the hidden representations, and then feed them into a Multi-Layer Perceptron (MLP). We model $p_\psi(\mathbf{z_y}|\mathbf{z_x},\mathbf{c})$ following a categorical distribution $\mathrm{Cat}(\boldsymbol{\pi})$, by computing the parameter $\boldsymbol{\pi}$ from $\mathbf{z_x}$ and the hidden representation of the context using another MLP.

Posterior Networks We use two conditional posterior networks $q_\phi(\mathbf{z_x}|\mathbf{x},\mathbf{c}), q_\phi(\mathbf{z_y}|\mathbf{z_x},\mathbf{y},\mathbf{c})$ to approximate true posterior distributions of latent variables for both question $\mathbf{x}$ and answer $\mathbf{y}$. We use two Bi-LSTM encoders to output the hidden representations of question and context given their word embeddings. Then, we feed the two hidden representations into MLP to obtain the parameters of Gaussian distribution, $\boldsymbol{\mu}'$ and $\boldsymbol{\sigma}'$ (upper right corner in Figure 1-(a)). We use the reparameterization trick (Kingma and Welling, 2014) to train the model with backpropagation since the stochastic sampling process $\mathbf{z_x} \sim q_\phi(\mathbf{z_x}|\mathbf{x},\mathbf{c})$ is nondifferentiable. We use another Bi-LSTM to encode the word embedding of answer-aware context into the hidden representation. Then, we feed the hidden representation and $\mathbf{z_x}$ into MLP to compute the parameters $\boldsymbol{\pi}'$ of categorical distribution (lower right corner in Figure 1-(a)). We use the categorical reparameterization trick with gumbel-softmax

(Maddison et al., 2017; Jang et al., 2017) to enable backpropagation through sampled discrete latent variables.

Answer Generation Networks Since we consider extractive QA, we can factorize $p_\theta(\mathbf{y}|\mathbf{z_y}, \mathbf{c})$ into $p_\theta(y_s|\mathbf{z_y}, \mathbf{c})$ and $p_\theta(y_e|\mathbf{z_y}, \mathbf{c})$, where y_s and y_e are the start and the end position of an answer span (highlighted words in Figure 1-(b)), respectively. To obtain MLE estimators for both, we first encode the context $\mathbf{c}$ into the contextualized word embedding of $\mathbf{E^c} = \{\mathbf{e}^\mathbf{c}_1, \ldots, \mathbf{e}^\mathbf{c}_M\}$ with the pre-trained BERT. We compute the final hidden representation of context and the latent variable $\mathbf{z_y}$ with a heuristic matching layer (Mou et al., 2016) and a Bi-LSTM:

$$\begin{aligned}\mathbf{f}_i &= [\mathbf{e}^\mathbf{c}_i;\ \mathbf{z_y};\ |\mathbf{e}^\mathbf{c}_i - \mathbf{z_y}|;\ \mathbf{e}^\mathbf{c}_i \odot \mathbf{z_y}]\\ \overrightarrow{\mathbf{h}}_i &= \overrightarrow{\text{LSTM}}([\mathbf{f}_i,\ \overrightarrow{\mathbf{h}}_{i-1}])\\ \overleftarrow{\mathbf{h}}_i &= \overleftarrow{\text{LSTM}}([\mathbf{f}_i,\ \overleftarrow{\mathbf{h}}_{i+1}])\\ \mathbf{H} &= [\ \overrightarrow{\mathbf{h}}_i;\ \overleftarrow{\mathbf{h}}_i\ ]_{i=1}^M\end{aligned}$$

where $\mathbf{z_y}$ is linearly transformed, and $\mathbf{H} \in \mathbb{R}^{d_\mathbf{y} \times M}$ is the final hidden representation. Then, we feed $\mathbf{H}$ into two separate linear layers to predict y_s and y_e.

Question Generation Networks We design the encoder-decoder architecture for our QG network by mainly adopting from our baselines (Zhao et al., 2018; Zhang and Bansal, 2019). For encoding, we use pre-trained BERT to encode the answer-specific context into the contextualized word embedding, and then use a two-layer Bi-LSTM to encode it into the hidden representation (in Figure 1-(c)). We apply a gated self-attention mechanism (Wang et al., 2017) to the hidden representation to better capture long-term dependencies within the context, to obtain a new hidden representation $\hat{\mathbf{H}} \in \mathbb{R}^{d_\mathbf{x} \times M}$.

The decoder is a two-layered LSTM which receives the latent variable $\mathbf{z_x}$ as an initial state. It uses an attention mechanism (Luong et al., 2015) to dynamically aggregate $\hat{\mathbf{H}}$ at each decoding step into a context vector of $\mathbf{s}_j$, using the j-th decoder hidden representation $\mathbf{d}_j \in \mathbb{R}^{d_\mathbf{x}}$ (in Figure 1-(c)). Then, we feed $\mathbf{d}_j$ and $\mathbf{s}_j$ into MLP with maxout activation (Goodfellow et al., 2013) to compute the final hidden representation $\hat{\mathbf{d}}_j$ as follows:

$$\begin{gathered}\mathbf{d}_0 = \mathbf{z_x},\ \ \mathbf{d}_j = \text{LSTM}([\mathbf{e}^\mathbf{x}_{j-1}, \mathbf{d}_{j-1}])\\ \mathbf{r}_j = \hat{\mathbf{H}}^T\mathbf{W^a}\mathbf{d}_j,\ \ \mathbf{a}_j = \text{softmax}(\mathbf{r}_j),\ \ \mathbf{s}_j = \hat{\mathbf{H}}\mathbf{a}_j\\ \hat{\mathbf{d}}_j = \text{MLP}([\ \mathbf{d}_j;\ \mathbf{s}_j\ ])\end{gathered}$$

where $\mathbf{z_x}$ is linearly transformed, and $\mathbf{e}^\mathbf{x}_j$ is the j-th question word embedding. The probability vector over the vocabulary is computed as $p(\mathbf{x}_j|\,\mathbf{x}_{<j}, \mathbf{z_x}, \mathbf{y}, \mathbf{c}) = \text{softmax}(\mathbf{W^e}\hat{\mathbf{d}}_j)$. We initialize the weight matrix $\mathbf{W^e}$ as the pretrained word embedding matrix and fix it during training. Further, we use the copy mechanism (Zhao et al., 2018), so that the model can directly copy tokens from the context. We also greedily decode questions to ensure that all stochasticity comes from the sampling of the latent variables.

3.2 Consistent QA Pair Generation with Mutual Information Maximization

One of the most important challenges of the QAG task is enforcing consistency between the generated question and its corresponding answer. They should be semantically consistent, such that it is possible to predict the answer given the question and the context. However, neural QG or QAG models often generate questions irrelevant to the context and the answer (Zhang and Bansal, 2019) due to the lack of the mechanism enforcing this consistency. We tackle this issue by maximizing the mutual information (MI) of a generated QA pair, assuming that an answerable QA pair will have high MI. Since an exact computation of MI is intractable, we use a neural approximation. While there exist many different approximations (Belghazi et al., 2018; Hjelm et al., 2019), we use the estimation proposed by Yeh and Chen (2019) based on Jensen-Shannon Divergence:

$$\begin{aligned}\text{MI}(X;Y) \geq\ & \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathbb{P}}[\log g(\mathbf{x},\mathbf{y})]\\ &+ \frac{1}{2}\mathbb{E}_{\tilde{\mathbf{x}},\mathbf{y}\sim\mathbb{N}}[\log(1 - g(\tilde{\mathbf{x}},\mathbf{y}))]\\ &+ \frac{1}{2}\mathbb{E}_{\mathbf{x},\tilde{\mathbf{y}}\sim\mathbb{N}}[\log(1 - g(\mathbf{x},\tilde{\mathbf{y}}))]\\ &=: \mathcal{L}_{\text{Info}}\end{aligned}$$

where $\mathbb{E}_\mathbb{P}$ and $\mathbb{E}_\mathbb{N}$ denote expectation over positive and negative examples. We generate negative examples by shuffling the QA pairs in the minibatch, such that a question is randomly associated with an answer. Intuitively, the function $g(\cdot)$ acts like a binary classifier that discriminates whether QA pair is from joint distribution or not. We empirically find that the following $g(\cdot)$ effectively achieves our goal of consistent QAG:

$$g(\mathbf{x},\mathbf{y}) = \text{sigmoid}(\overline{\mathbf{x}}^T\mathbf{W}\overline{\mathbf{y}})$$

where $\overline{\mathbf{x}} = \frac{1}{N}\sum_i \hat{\mathbf{d}}_i$ and $\overline{\mathbf{y}} = \frac{1}{L}\sum_j \hat{\mathbf{h}}_j$ are summarized representations of question and answer, respectively. Combined with the ELBO, the final objective of our Info-HCVAE is as follows:

$$\max_{\Theta} \mathcal{L}_{\text{HCVAE}} + \lambda \mathcal{L}_{\text{Info}}$$

where Θ includes all the parameters of ϕ, ψ, θ and $\mathbf{W}$, and λ controls the effect of MI maximization. In all experiments, we always set the λ as 1.

4 Experiment

4.1 Dataset

Stanford Question Answering Dataset v1.1 (SQuAD) (Rajpurkar et al., 2016). This is a reading comprehension dataset consisting of questions obtained from crowdsourcing on a set of Wikipedia articles, where the answer to every question is a segment of text or a span from the corresponding reading passage. We use the same split used in Zhang and Bansal (2019) for the fair comparison.

Natural Questions (NQ) (Kwiatkowski et al., 2019). This dataset contains realistic questions from actual user queries to a search engine, using Wikipedia articles as context. We adapt the dataset provided from MRQA shared task (Fisch et al., 2019) and convert it into the extractive QA format. We split the original validation set in half, to use as validation and test for our experiments.

TriviaQA (Joshi et al., 2017). This is a reading comprehension dataset containing question-answer-evidence triples. The QA pairs and the evidence (contexts) documents are authored and uploaded by Trivia enthusiasts. Again, we only choose QA pairs of which answers are span of contexts.

HarvestingQA [2] This dataset contains top-ranking 10K Wikipedia articles and 1M synthetic QA pairs generated from them, by the answer span extraction and QG system proposed in (Du and Cardie, 2018). We use this dataset for semi-supervised learning.

4.2 Experimental Setups

Implementation Details In all experiments, we use BERT-base ($d = 768$) (Devlin et al., 2019) as the QA model, setting most of the hyperparameters as described in the original paper. For both HCVAE and Info-HCVAE, we set the hidden dimensionality of the Bi-LSTM to 300 for posterior, prior, and answer generation networks, and use the dimensionality of 450 and 900 for the encoder and the decoder of the question generation network. We set the dimensionality of $\mathbf{z_x}$ as 50, and define $\mathbf{z_y}$ to be set of 10-way categorical variables $\mathbf{z_y} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{20}\}$. For training the QA model, we fine-tune the model for 2 epochs. We train both the QA model and Info-HCVAE with Adam optimizer (Kingma and Ba, 2015) with the batch size of 32 and the initial learning rate of $5 \cdot 10^{-5}$ and 10^{-3} respectively. For semi-supervised learning, we first pre-train BERT on the synthetic data for 2 epochs and fine-tune it on the GT dataset for 2 epochs. To prevent *posterior collapse*, we multiply 0.1 to the KL divergence terms of question and answer (Higgins et al., 2017). For more details of the datasets and experimental setup, please see **Appendix** C.

Baselines We experiment two variants of our model against several baselines:

1. **Harvest-QG**: An attention-based neural QG model with a neural answer extraction system (Du and Cardie, 2018).
2. **Maxout-QG**: A neural QG model based on maxout copy mechanism with a gated self-attetion (Zhao et al., 2018), which uses BERT as the word embedding as suggested by Zhang and Bansal (2019).
3. **Semantic-QG**: A neural QG model based on Maxout-QG with semantic-enhanced reinforcement learning (Zhang and Bansal, 2019).
4. **HCVAE**: Our HCVAE model without the Info-Max regularizer.
5. **Info-HCVAE**: Our full model with the InfoMax regularizer.

For the baselines, we use the same answer spans extracted by the answer extraction system (Du and Cardie, 2018).

4.3 Quantitative Analysis

QAE and R-QAE One of crucial challenges with generative models is a lack of a good quantitative evaluation metric. We adopt **QA**-based **E**valuation (QAE) metric proposed by Zhang and Bansal (2019) to measure the quality of QA pair. QAE is obtained by first training the QA model on the synthetic data, and then evaluating the QA model with human annotated test data. However, QAE only measures how well the distribution of synthetic QA pairs matches the distribution of GT QA pairs, and does not consider the diversity of QA pairs. Thus, we propose **R**everse **QA**-based **E**valuation (**R-QAE**), which is the accuracy of the QA model trained on the human-annotated QA pairs, evaluated on the generated QA pairs. If the synthetic

[2] https://github.com/xinyadu/harvestingQA

Method	QAE (↑)	R-QAE (↓)
SQuAD (EM/F1)		
Harvesting-QG	55.11/66.40	64.77/78.85
Maxout-QG	56.08/67.50	62.49/78.24
Semantic-QG	60.49/71.81	74.23/88.54
HCVAE	69.46/80.79	**37.57**/61.24
Info-HCVAE	**71.18/81.51**	38.80/**60.73**
Natural Questions (EM/F1)		
Harvesting-QG	27.91/41.23	49.89/70.01
Maxout-QG	30.98/44.96	49.96/70.03
Semantic-QG	30.59/45.29	58.42/79.23
HCVAE	31.45/46.77	32.78/55.12
Info-HCVAE	**37.18/51.46**	**29.39/53.04**
TriviaQA (EM/F1)		
Harvesting-QG	21.32/30.21	29.75/47.73
Maxout-QG	24.58/34.32	31.56/49.92
Semantic-QG	27.54/38.25	37.45/58.15
HCVAE	30.20/40.88	34.41/48.16
Info-HCVAE	**35.45/44.11**	**21.65/37.65**

Table 2: QAE and R-QAE results on three datasets. All results are the performances on our test set.

Harvest -QG	Maxout -QG	Semantic -QG	HCVAE	Info-HCVAE
111.74	114.58	112.94	113.89	**117.41**

Table 3: The results of mutual information estimation. The results are based on QA pairs generated from H×10%.

data covers larger distribution than the human annotated training data, R-QAE will be lower. However, note that having a low R-QAE is only meaningful when the QAE is high enough since trivially invalid questions may also yield low R-QAE.

Results We compare HCVAE and Info-HCVAE with the baseline models on SQuAD, NQ, and TriviaQA. We use 10% of Wikipedia paragraphs from HarvestingQA (Du and Cardie, 2018) for evaluation. Table 2 shows that both HCVAE and Info-HCVAE significantly outperforms all baselines by large margin in QAE on all three datasets, while obtaining significantly lower R-QAE, which shows that our model generated both high-quality and diverse QA pairs from the given context. Moreover, Info-HCVAE largely outperforms HCVAE, which demonstrates the effectiveness of our InfoMax regularizer for enforcing QA-pair consistency.

Figure 3 shows the accuracy as a function of number of QA pairs. Our Info-HCVAE outperform all baselines by large margins using orders of magnitude smaller number of QA pairs. For example, Info-HCVAE achieves 61.38 points using 12K QA pairs, outperforming Semantic-QG that use 10 times larger number of QA pairs. We also report

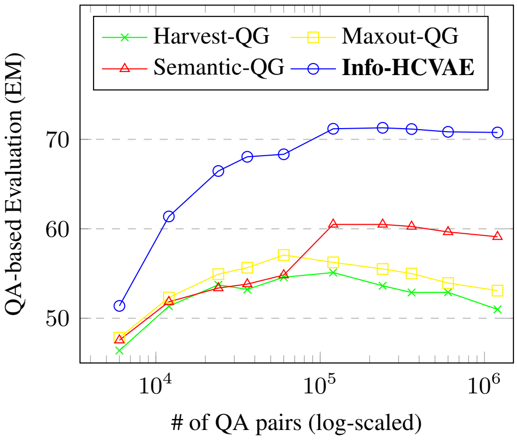


Figure 3: QAE vs. # of QA pairs (log-scaled) on SQuAD.

Method	QAE (↑)	R-QAE (↓)
Baseline	56.08/67.50	62.49/78.24
+Q-latent	58.66/70.54	40.00/62.02
+A-latent	69.46/80.79	**37.57**/61.24
+InfoMax	**71.18/81.51**	38.80/**60.73**

Table 4: QAE and R-QAE results of the ablation study on SQuAD dataset. All the results are the performances on our test set.

the score of $\overline{\mathbf{x}}^T\mathbf{W}\overline{\mathbf{y}}$ as an approximate estimation of mutual information (MI) between QA pairs generated by each method in Table 3; our Info-HCVAE yields the largest value of MI estimation.

Ablation Study We further perform an ablation study to see the effect of each model component. We start with the model without any latent variables, which is essentially a deterministic Seq2Seq model (denoted as Baseline in Table 4). Then, we add in the question latent variable (+Q-latent) and then the answer latent variable (+A-latent), to see the effect of probabilistic latent variable modeling and hierarchical modeling respectively. The results in Table 4 shows that both are essential for improving both the quality (QAE) and diversity (R-QAE) of the generated QA pairs. Finally, adding in the InfoMax regularization (+InfoMax) further improves the performance by enhancing the consistency of the generated QA pairs.

4.4 Qualitative Analysis

Human Evaluation As a qualitative analysis, we first conduct a pairwise human evaluation of the QA pairs generated by our Info-HCVAE and Maxout-QG on 100 randomly selected paragraphs. Specifically, 20 human judges performed blind quality assessment of two sets of QA pairs that are presented in a random order, each of which contained two to five QA pairs. Each set of QA pairs is evalu-

Method	**Diversity**	**Consistency**	**Overall**
Baseline	26%	34%	30%
Ours	**47%**	**50%**	**52%**
Tie	27%	16%	18%

Table 5: The results of human judgement in terms of diversity, consistency, and overall quality on the generated QA pairs.

Paragraph	The scotland act 1998 which was passed by and given royal assent by queen Elizabeth ii on 19 november 1998, governs functions and role of the scottish parliament and delimits its legislative competence . . .
GT	what act sets forth the functions of the scottish parliament?
O-1	which act was passed in 1998?
O-2	which act governs role of the scottish parliament?
O-3	which act was passed by queen Elizabeth ii?
O-4	which act gave the scottish parliament the responsibility to determine its legislative policy?

Table 6: Examples of *one-to-many* mapping of our Info-HCVAE. The answer is highlighted by pink. **GT** denotes the ground-truth question. **O-** denotes questions generated by Info-HCVAE.

Paragraph ... Atop the main building' s gold dome is a golden statue of the virgin mary. ... Next to the main building is the basilica of the sacred heart. Immediately behind the basilica is the grotto, ... a marian place of prayer and reflection. ... At the end of the main drive ..., is a simple, modern stone statue of mary.	
Ori1	**Q** what is the grotto at notre dame? **A** a marian place of prayer and reflection
↑	**Q** *where is the grotto at?* **A** *a marian place of prayer and reflection*
	Q *what place is behind the basilica of prayer?* **A** *grotto*
Gen	**Q** *what is next to the main building at notre dame?* **A** *the basilica of the sacred heart*
↓	**Q** *what is at the end of the main drive?* **A** *stone statue of mary*
Ori2	**Q** what sits on top of the main building at notre dame? **A** a golden statue of the virgin mary

Table 7: QA pairs generated by interpolating between two latent codes encoded by our posterior networks. **Ori1** and **Ori2** are from training set of SQuAD.

ated in terms of the overall quality, diversity, and consistency between the generated QA pairs and the context. The results in Table 5 show that the QA pairs generated by our Info-HCVAE is evaluated to be more diverse and consistent, compared to ones generated by the baseline models.

One-to-Many QG To show that our Info-HCVAE can effectively tackle *one-to-many* mapping problem for question generation, we qualitatively analyze the generated questions for given a context and an answer from the SQuAD validation set. Specifically, we sample the question latent variables multiple times using the question prior network $p_\psi(\mathbf{z_x} \mid \mathbf{c})$, and then feed them to question generation networks $p_\theta(\mathbf{x} \mid \mathbf{z_x}, \mathbf{y}, \mathbf{c})$ with the answer. The example in Table 6 shows that our Info-HCVAE generates diverse and semantically consistent questions given an answer. We provide more qualitative examples in **Appendix** D.

Latent Space Interpolation To examine if Info-HCVAE learns meaningful latent space of QA pairs, we qualitatively analyze the QA pairs generated by interpolating between two latent codes of it on SQuAD training set. We first encode $\mathbf{z_x}$ from two QA pairs using posterior networks of $q_\phi(\mathbf{z_x}|\mathbf{x}, \mathbf{c})$, and then sample $\mathbf{z_y}$ from interpolated values of $\mathbf{z_x}$ using prior networks $p_\psi(\mathbf{z_y}|\mathbf{z_x}, \mathbf{c})$ to generate corresponding QA pairs. Table 7 shows that the semantic of the QA pairs generated smoothly transit from one latent to another with high diversity and consistency. We provide more qualitative examples in **Appendix** D.

4.5 Semi-supervised QA

We now validate our model in a semi-supervised setting, where the model uses both the ground truth labels and the generated labels to solve the QA task, to see whether the generated QA pairs help improve the performance of a QA model in a conventional setting. Since such synthetic datasets consisting of generated QA pairs may inevitably contain some noise (Zhang and Bansal, 2019; Dong et al., 2019; Alberti et al., 2019), we further refine the QA pairs by using the heuristic suggested by Dong et al. (2019), to replace the generated answers whose F1 score to the prediction of the QA model trained on the human annotated data is lower than a set threshold. We select the threshold of 40.0 for the QA pair refinement model via cross-validation on the SQuAD dataset, and used it for the experiments. Please see **Appendix** C for more details.

SQuAD We first perform semi-supervised QA experiments on SQuAD using the synthetic QA pairs generated by our model. For the contexts, we use both the paragraphs in the original SQuAD (S) dataset, and the new paragraphs in the HarvestingQA dataset (H). Using Info-HCVAE, we generate 10 different QA pairs by sampling from the latent spaces (denoted as S$\times$10). For the baseline, we use Semantic-QG (Zhang and Bansal, 2019) with the beam search size of 10 to obtain the same number of QA pairs. We also generate new QA pairs

Data	EM	F1
SQuAD	80.25	88.23
Semantic-QG (baseline)		
+S×10	81.20 (+0.95)	88.36 (+0.13)
+H×100%	81.03 (+0.78)	88.79 (+0.56)
+S×10 + H×100%	81.44 (+1.19)	88.72 (+0.49)
Info-HCVAE (ours)		
+S×10	82.09 (+1.84)	89.11 (+0.88)
+H×10%	81.37 (+1.12)	88.85 (+0.62)
+H×20%	81.68 (+1.43)	89.06 (+0.93)
+H×30%	81.76 (+1.51)	89.12 (+0.89)
+H×50%	82.17 (+1.92)	89.38 (+1.15)
+H×100%	**82.37 (+2.12)**	89.63 (+1.40)
+S×10 + H×100%	82.19 (+1.94)	**89.84 (+1.59)**

Table 8: The results of semi-supervised QA experiments on SQuAD. All the results are the performances on our test set.

using different portions of paragraphs provided in HarvestingQA (denoted as H×10%-H×100%), by sampling one latent variable per context. Table 8 shows that our framework improves the accuracy of the BERT-base model by 2.12 (EM) and 1.59 (F1) points, significantly outperforming Semantic-QG.

NQ and TriviaQA Our model is most useful when we do not have any labeled data for a target dataset. To show how well our QAG model performs in such a setting, we train the QA model using only the QA pairs generated by our model trained on SQuAD and test it on the target datasets (NQ and TriviaQA). We generate multiple QA pairs from each context of the target dataset, sampling from the latent space one to ten times (denoted by N×1-10 or T×1-10 in Table 9). Then, we fine-tune the QA model pretrained on the SQuAD dataset with the generated QA pairs from the two datasets. Table 9 shows that as we augment training data with larger number of synthetic QA pairs, the performance of the QA model significantly increases, significantly outperforming the QA model trained on SQuAD only. Yet, models trained with our QAG still largely underperform models trained with human labels, due to the distributional discrepancy between the source and the target dataset.

Data	EM	F1
Natural Questions		
SQuAD	42.77	57.29
+N×1	46.70 (+3.94)	61.08 (+3.79)
+N×2	46.95 (+4.19)	61.34 (+4.05)
+N×3	47.73 (+4.96)	61.98 (+4.69)
+N×5	48.19 (+5.42)	62.21 (+4.92)
+N×10	**48.44 (+5.67)**	**62.69 (+5.40)**
NQ	61.65	73.91
TriviaQA		
SQuAD	48.96	57.98
+T×1	49.65 (+0.69)	59.13 (+1.21)
+T×2	50.01 (+1.05)	59.08 (+1.10)
+T×3	49.71 (+0.75)	**59.49 (+1.51)**
+T×5	**50.14 (+1.18)**	59.21 (+1.23)
+T×10	49.65 (+0.69)	59.20 (+1.22)
Trivia	64.55	70.42

Table 9: The result of semi-supervised QA experiments on Natural Questions and TriviaQA dataset. All results are the performance on our test set.

5 Conclusion

We proposed a novel probabilistic generative framework for generating diverse and consistent question-answer (QA) pairs from given texts. Specifically, our model learns the joint distribution of question and answer given context with a hierarchically conditional variational autoencoder, while enforcing consistency between generated QA pairs by maximizing their mutual information with a novel InfoMax regularizer. To our knowledge, ours is the first successful probabilistic QAG model. We evaluated the QAG performance of our model by the accuracy of the BERT-base QA model trained using the generated questions on multiple datasets, on which it largely outperformed the state-of-the-art QAG baseline (+6.59-10.69 in EM), even with a smaller number of QA pairs. We further validated our model for semi-supervised QA, where it improved the performance of the BERT-base QA model on the SQuAD by 2.12 in EM, significantly outperforming the state-of-the-art model. As future work, we plan to extend our QAG model to a meta-learning framework, for generalization over diverse datasets.

Acknowledgements

This work was supported by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01410, Research Development of Question Generation for Deep Learning based Semantic Search Domain Extension, No.2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion, No.2019- 0-00075, and Artificial Intelligence Graduate School Program (KAIST)).

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. 2018. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems, NIPS 2018*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NACCL-HLT, 2018*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems, NeurIPS, 2019*.

Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational autoregressive decoder for neural response generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. EMNLP 2018*.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *EMNLP 2019 MRQA Workshop*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems, NIPS 2014*.

Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. In *Proceedings of the 30th International Conference on International Conference on Machine, ICML 2013*.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NACCL-HLT 2010*.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations, ICLR 2019*.

Tom Hosking and Sebastian Riedel. 2019. Evaluating rewards for question generation models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of* *the Association for Computational Linguistics, ACL 2017*.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Diederik P. Kingma and Max Welling. 2014. Autoencoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018. A framework for automatic question generation from text using deep reinforcement learning. *CoRR*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics, TACL 2019*.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational, ACL 2015*.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*.

David Lindberg, Fred Popowich, John C. Nesbit, and Philip H. Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation, ENLG 2013*.

Bang Liu, Mingjun Zhao, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. 2019. Learning to generate questions by learningwhat not to generate. In *The World Wide Web Conference, WWW 2019*.

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, EMNLP 2015*.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017*.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

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

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NACCL-HLT 2018*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*.

Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. Are training samples correlated? learning to generate dialogue responses with multiple references. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*.

Mrinmaya Sachan and Eric P. Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence 2018*.

Linfeng Song, Zhiguo Wang, and Wael Hamza. 2017. A unified query-based generative model for question generation and question answering. *CoRR*.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence 2018*.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, EMNLP 2018*.

Duyu Tang, Nan Duan, Tao Qin, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *CoRR*.

Duyu Tang, Nan Duan, Zhao Yan, Zhirui Zhang, Yibo Sun, Shujie Liu, Yuanhua Lv, and Ming Zhou. 2018. Learning to collaborate for question answering and asking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*.

Kaichun Yao, Libo Zhang, Tiejian Luo, Lili Tao, and Yanjun Wu. 2018. Teaching machines to ask questions. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*.

Yi-Ting Yeh and Yun-Nung Chen. 2019. Qainfomax: Learning robust question answering system by mutual information maximization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*.

Xingdi Yuan, Tong Wang, Çaglar Gülçehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017*.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. *CoRR*.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*.

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

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

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017*.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*.

Appendix

A Derivation of Variational Lower Bound

Theorem. *If we assume conditional independence of* $\mathbf{y}$ *and* $\mathbf{z}_{\mathbf{x}}$*, i.e.,* $p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{y}}, \mathbf{c}) = p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c})$*,* $\log p_\theta(\mathbf{x}, \mathbf{y}|\mathbf{c}) \geq \mathcal{L}_{HCVAE}$

Proof.

$$
\begin{aligned}
&\log p_\theta(\mathbf{x}, \mathbf{y}|\mathbf{c}) \\
&= \log \int_{\mathbf{z}_{\mathbf{x}}} \sum_{\mathbf{z}_{\mathbf{y}}} p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c}) \cdot \\
&\quad p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{y}}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c}) d_{\mathbf{z}_{\mathbf{x}}} \\
&= \log \int_{\mathbf{z}_{\mathbf{x}}} p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c}) \frac{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})} \cdot \\
&\quad \sum_{\mathbf{z}_{\mathbf{y}}} p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c}) \frac{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} d_{\mathbf{z}_{\mathbf{x}}} \\
&= \log \int_{\mathbf{z}_{\mathbf{x}}} p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c}) \frac{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})} \\
&\quad \cdot \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \left[\frac{p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \right] d_{\mathbf{z}_{\mathbf{x}}} \\
&= \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})} \{ \frac{p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})} \cdot \\
&\quad \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \left[\frac{p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \right] \} \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})} \{ \log \frac{p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})} + \\
&\quad \log \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \left[\frac{p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \right] \} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})} [\log p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})] \\
&\quad - \mathrm{D}_{\mathrm{KL}}[q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c}) || p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c})] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})} \{ \\
&\quad \log \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \left[\frac{p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c}) p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c})}{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} \right] \} \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})} [\log p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})] \\
&\quad - \mathrm{D}_{\mathrm{KL}}[q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c}) || p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c})] \\
&\quad + \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})} \{ \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} [\log p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c})] \\
&\quad - \mathrm{D}_{\mathrm{KL}}[q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c}) || p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c})] \} \\
&\approx \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c})} [\log p_\theta(\mathbf{x}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})] \\
&\quad - \mathrm{D}_{\mathrm{KL}}[q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x}, \mathbf{c}) || p_\psi(\mathbf{z}_{\mathbf{x}}|\mathbf{c})] \\
&\quad + \mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c})} [\log p_\theta(\mathbf{y}|\mathbf{z}_{\mathbf{y}}, \mathbf{c})] \\
&\quad - \mathrm{D}_{\mathrm{KL}}[q_\phi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{y}, \mathbf{c}) || p_\psi(\mathbf{z}_{\mathbf{y}}|\mathbf{z}_{\mathbf{x}}, \mathbf{c})]
\end{aligned}
$$

□

B Datatset

The statistics and the data resource are summarized in Table 10.

SQuAD We tokenize questions and contexts with WordPiece tokenizer from BERT. To fairly compare our proposed methods with the existing semi-supervised QA, we follow Zhang and Bansal (2019)'s split, which divides original development set from SQuAD v1.1 (Rajpurkar et al., 2016) into new validation set and test set. We adopt most of the codes from Wolf et al. (2019) for preprocessing data, training, and evaluating the BERT-base QA model.

Natural Questions Other than the original Natural Questions (Kwiatkowski et al., 2019) dataset, we use subset of the dataset provided by MRQA shared task (Fisch et al., 2019) for extractive QA. As semi-supervised setting with SQuAD, we split the validation set provided from MRQA into half for validation set and the others for test set. All the tokens from question and context are tokenized with WordPiece tokenizer from BERT. We generate QA pairs from context not containing html tag, and evaluate QA model with the official MRQA evaluation scripts.

TriviaQA For TriviaQA (Joshi et al., 2017), we also use the training set from MRQA shared task, and divide the development set from MRQA into half for validation set and the other for test set. All the tokens from question and context are tokenized with WordPiece tokenizer from BERT. For evaluation, we follow the MRQA's official evaluation procedure.

HarvestingQA[3] We use paragraphs from HarvestingQA dastaset (Du and Cardie, 2018) to generate QA pairs for QA-based Evaluation (QAE) and Reverse QA-based Evaluation (R-QAE). For the baseline QG models such as Maxout-QG and Semantic-QG, we use the same answer spans from the dataset. For the experiments of Maxout-QG baseline, we train the model and generate new questions from the context and answer, while the questions generated by Semantic-QG are provided by the authors (Zhang and Bansal, 2019).

C Training Details

Maxout-QG We use Adam (Kingma and Ba, 2015) optimizer with the batch size of 64 and set the initial learning rate of 10^{-3}. We always set the

[3] https://github.com/xinyadu/harvestingQA

Datasets	Train (#)	Valid (#)	Source
SQuAD	86,588	10,507	Crowd-sourced questions from Wikipedia paragraph
Natural Questions	104,071	12,836	Questions from actual userfor searching Wikipedia paragraph
TriviaQA	74,160	7,785	Question and answer pairs authored by trivia enthusaists from the Web
HarvestQA	1,259,691	-	Generated by neural networks from top-ranking 10,000 Wikipedia articles

Table 10: The statistics and the data source of SQuAD, Natural Questions, TriviaQA, and HarvestingQA.

Replace	EM	F1
F1 $\leq$ 0.0	82.4	89.39
F1 $\leq$ 20.0	83.11	89.65
F1 $\leq$ 40.0	**83.32**	**89.79**
F1 $\leq$ 60.0	83.20	89.78
F1 $\leq$ 80.0	83.09	89.75

Table 11: The effect of F1-based replacement strategy in semi-supervised setting of SQuAD+H×100%. All results are the performance on validation set of Zhang and Bansal (2019).

beam size of 10 for decoding. We also evaluate the Maxout-QG model on our SQuAD validation set with BLEU4 (Papineni et al., 2002), and get 15.68 points.

Selection of Threshold for Replacement As mentioned in our paper, we use the threshold of 40.0 selected via cross-validation of the QA model performance, using both the full SQuAD and HarvestingQA dataset for QAG. The detailed selection processes are as follows: 1) train QA model on only human annotated data, 2) compute F1 score of generated QA pairs, and 3) if the F1 score is lower than the threshold, replace the generated answer with the prediction of QA model. We investigate the optimal value of threshold among $[20.0, 40.0, 60.0, 80.0]$ using our validation set of SQuAD. Table 11 shows the results of cross-validation on the validation set. The optimal value of 40.0 is used for semi-supervised experiments on Natural Questions and TriviaQA. For fully unlabeled semi-supervised experiments on Natural Questions and TriviaQA, the QA model is only trained on SQuAD and used to replace the synthetic QA pairs (denoted in our paper as N×1-10, T×1-10).

Semi-supervised learning For the semi-supervised learning experiment on SQuAD, we follow Zhang and Bansal (2019)'s split for a fair comparison. Specifically, we receive the unique IDs for QA pairs from the authors and use exactly the same validation and test set as theirs. For the Natural Questions and TriviaQA experiments, we use our own split as mentioned in the above. We generate QA pairs from the paragraphs of Wikipedia extracted by Du and Cardie (2018) and train BERT-base QA model with the synthetic data for two epochs. Then we further train the model with human-annotated training data for two more epochs. The catastrophic forgetting reported in Zhang and Bansal (2019) does not occur in our cases. We use Adam optimizer (Kingma and Ba, 2015) with batch size 32 and follow the learning rate scheduling as described in (Devlin et al., 2019) with initial learning rate $2 \cdot 10^{-5}$ and $3 \cdot 10^{-5}$ for synthetic and human annotated data, respectively.

D Qualitative Examples

The qualitative examples in Table 12, 13, 14 are shown in the next page.

Paragraph-1 Near Tamins-Reichenau the Anterior Rhine and the Posterior Rhine join and form the Rhine. . . . This section is nearly 86km long, and descends from a height of 599m to 396m. It flows through a wide glacial alpine valley known as the Rhine Valley (German: Rheintal). Near Sargans a natural dam, only a few metres high, . . . The Alpine Rhine begins in the most western part of the Swiss canton of Graubünden, . . .
Q-1: how long is the rhine? **A-1**: 86km long
Q-2: how large is the dam? **A-2**: a few metres high
Q-3: where does the anterior rhine and the posterior rhine join the rhine? **A-3**: Tamins-Reichneau
Q-4: what type of valley does the rhine flows through? **A-4**: glacial alpine
Q-5: what is the rhine valley in german? **A-5**: Rheintal
Q-6: where deos the alpine rhine begin? **A-7**: Swiss canton of Graubünden
Paragraph-2 Victoria is the centre of dairy farming in Australia. It is home to 60% of Australia's 3 million dairy cattle and produces nearly two-thirds of the nation's milk, almost 6.4 billion litres. The state also has 2.4 million beef cattle, with more than 2.2 million cattle and calves slaughtered each year. In 2003–04, Victorian commercial fishing crews and aquaculture industry produced 11,634 tonnes of seafood valued at nearly $109 million. . . .
Q-1: what industry produced 11,63 million tonnes of seafood in 2003-04 ? **A-1**: aquaculture
Q-2: what type of cattle is consumed in Victoria? **A-2**: beef
Q-3: in what year did victorian commercial fishing and aquaculture industry produce a large amount of seafood? **A-3**: 2003–04
Q-4: how many cattle and calves each year are slaughtered annually? **A-4**: 2.2 million
Q-5: how much of the nation's milk is produced by the dairy? **A-5**: two-thrids
Paragraph-3 A teacher's role may vary among cultures. Teachers may provide instruction in literacy and numeracy, craftsmanship or vocational training, the arts, religion, civics, community roles, or life skills.
Q-1: what do a teacher's role vary? **A-1**: culture
Q-2: what do teachers provide instruction in? **A-2**: vocational training
Q-3: what is one thing a teacher may provide instruction for? **A-3**: community roles
Q-4: what is one of the skills that teachers provide in? **A-4**: life skills

Table 12: Examples of QA pairs generated by our Info-HCVAE. We sample multiple latent variables from $p_\psi(\cdot)$, and feed them to generation networks. All the paragraphs are from validation set of SQuAD.

Paragraph-1	Super bowl 50 was an american football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24 – 10 to earn their third super bowl title. . . .
GT	which NFL team represented the AFC at super bowl 50?
Ours-1	what team did the American Football Conference represent?
Ours-2	who won the 2015 American Football Conference?
Ours-3	which team defeated the carolina panthers?
Ours-4	who defeated the panthers in 2015?
Ours-5	what team defeated the carolina panthers in the 2015 season?
Ours-6	who was the champion of the American Football League in the 2015 season?
Ours-7	what team won the 2015 American Football Conference?
Paragraph-2	. . . Some clergy offer healing services, while exorcism is an occasional practice by some clergy in the united methodist church in Africa. . . .
GT	in what country does some clergy in the umc occasionally practice exorcism?
Ours-1	in what country do some clergy in the united methodist church take place?
Ours-2	in what country is exorcism practice an occasional practice?
Ours-3	use of exorcism is an occasional practice in what country?
Ours-4	is exorcism usually an occasional practice in what country?
Paragraph-3	. . . , the city was the subject of a song , “walking into fresno” , written by hall of fame guitarist Bill Aken . . .
GT	who wrote “walking in fresno”?
Ours-1	who wrote “walking into fresno”?
Ours-2	“walking into fresno” was written by whom?
Ours-3	the song “walking into fresno” was written by whom?

Table 13: Examples of *one-to-many* mapping of our Info-HCVAE. Answers are highlighted by pink. We sample multiple question latent variables from $p_\psi(\mathbf{z}_\mathbf{x} \mid \mathbf{c})$, and feed them to question generation networks with a fixed answer. **GT** denotes ground-truth question, and **Seq2Seq** denotes question generated by Maxout-QG. All the paragraphs, ground truth questions, and answers are from validation set of SQuAD.

Paragraph-1 Notre Dame is known for its competitive admissions, with the incoming class enrolling in fall 2015 admitting 3,577 from a pool of 18,156 (19.7%). The academic profile of the enrolled class continues to rate among the top 10 to 15 in the nation for national research universities. . . . 1,400 of the 3,577 (39.1%) were admitted under the early action plan.	
Ori1	**Q** where does notre dame rank in terms of academic profile among research universities in the us? **A** the top 10 to 15 in the nation
↑	**Q** *where does the academic profile of notre dame rank?* **A** *the top 10 to 15*
Gen	**Q** *what was the rate of the incoming class enrolling in the fall of 2015?* **A** *3,577 from a pool of 18,156 (19.7%)*
↓	**Q** *how many students attended notre dame?* **A** *3,577*
Ori2	**Q** what percentage of students at notre dame participated in the early action program? **A** 39.1%
Paragraph-2 . . . begun as a one-page journal in September 1876, the scholastic magazine is issued twice monthly and . . . In 1987, when some students believed that the observer began to show a . . . In spring 2008 an undergraduate journal for political science research, beyond politics, made its debut.	
Ori1	**Q** when did the scholastic magazine of notre dame begin publishing? **A** september 1876
↑	**Q** *when was the scholastic magazine published?* **A** *1876*
Gen	**Q** *in what year did notre dame get its liberal newspaper?* **A** *1987*
↓	**Q** *how often is the scholastic magazine published ?* **A** *twice*
Ori2	**Q** in what year did notre dame begin its undergraduate journal ? **A** 2008
Paragraph-3 As at most other universities, notre dame's students run a number of news media outlets. The nine student - run outlets include . . . , and several magazines and journals. . . . . the dome yearbook is published annually. . . .	
Ori1	**Q** what is the daily student paper at notre dame called? **A** the observer
↑	**Q** *how many student media outlets are there at notre dame?* **A** *nine student - run outlets include three*
Gen	**Q** *what type of media is the student paper at notre dame?* **A** *a number of news media*
↓	**Q** *how often is the dome published?* **A** *annually*
	Q *how many magazines are published at notre dame ?* **A** *several*
Ori2	**Q** how many student news papers are found at notre dame ? **A** three

Table 14: QA pairs generated by interpolating between two latent codes encoded by our posterior networks. **Ori1** and **Ori2** are from training set of SQuAD.