On the Generation of Medical Question-Answer Pairs*

Sheng Shen[1], Yaliang Li[2], Nan Du[2], Xian Wu[2], Yusheng Xie[2], Shen Ge[2], Tao Yang[2], Kai Wang[2], Xingzheng Liang[3], and Wei Fan[2]
[1]University of California, Berkeley, CA USA
[2]Tencent Medical AI Lab, CA USA
[3]Zhejiang University, Hangzhou China
email: {ndu, davidwfan}@tencent.com

Abstract

Question answering (QA) has achieved promising progress recently. However, answering a question in real-world scenarios like the medical domain is still challenging, due to the requirement of external knowledge and the insufficient of high-quality training data. In the light of these challenges, we study the task of generating medical QA pairs in this paper. With the insight that each medical question can be considered as a sample from the latent distribution conditioned on the corresponding answer, we propose an automated medical QA pair generation framework, consisting of an unsupervised key phrase detector that explores unstructured material for validity, and a generator that involves multi-pass decoder to integrate with structural knowledge for diversity. Series of experiments have been conducted on a real-world dataset collected from the National Medical Licensing Examination of China. Both automatic evaluation and human annotation demonstrate the effectiveness of the proposed method. Further investigation shows that, by incorporating the generated QA pairs for training, significant improvement in terms of accuracy can be achieved for the examination QA system.

Introduction

Due to the remarkable breakthrough of deep learning and natural language processing, question answering (QA) has gained increasing popularity in the past few years. Among its broad application domains, medical QA is one of the most appealing real-world application scenarios: People tend to consult others about health-related issues on online-community, which is more efficient than visiting doctors.

Although QA systems with deep learning methods have achieved good performance, medical QA confronts more difficulties than other domains. First, medical QA system requires highly accurate answers, and thus external and professional knowledge gathered from various sources are needed. Second, the number of available high-quality medical QA pairs is limited, as the labeling process by medical experts is time-consuming and expensive. Therefore, the performance of medical QA system is further constrained by the paucity of high-quality QA pairs since it can hardly learn a good model from limited training data.

To tackle these difficulties, the generation of medical QA pairs plays an indispensable role. By the automatic generation of high-quality medical QA pairs, external and professional knowledge can be incorporated, and the size of training data can be augmented. Therefore, we study this important task of medical QA pair generation in this paper. To be more specific, we assume that each medical answer corresponds to a distribution of valid questions, which should be constrained on external medical knowledge. Following this assumption, with more high-quality QA pairs generated based on the same knowledge as original QA pairs, the latent distribution of available medical QA pairs can be supplemented and thus medical QA system could learn unbiased model easily.

However, the generation of new medical QA pairs based on original ones is challenging: It is hard to maintain the diversity and the validity of generated question-answer pairs simultaneously. Existing question-answer pair generation method (Duan et al. 2017) only focused on the word-level diversity, and it may generate similar question-answer pairs to the original ones, which are valid but useless to augment the original medical QA data as these generated similar QA pairs do not introduce any new knowledge. On the other hand, if more diversity is promoted, such as focusing on sentence-level diversity, validity might not be guaranteed.

To ensure the validity of the generated medial QA pairs, we propose a retrieval and matching method to detect the key information of QA pairs in an unsupervised way from unstructured text materials such as patients' medical records, textbooks, and research articles.

To explore the diversity of the generated medical QA pairs while maintaining validity, we propose two mechanisms to incorporate structured and unstructured knowledge for the generation. We first explore global phrase level diversity and validity with a hierarchical Conditional Variational Autoencoder (CVAE) based framework, which models phrase level relationships in the original medical QA pairs, and generates the new pairs without breaking these relationships. We then propose a multi-pass decoder, in which all the local components (type, entities in each phrase) are coupled together and are jointly optimized in an end-to-end fashion.

In order to demonstrate the effectiveness of the proposed generation method, we evaluate the quality of generated medical QA pairs with both human annotation and several

*This work was done when the authors Sheng Shen and Xingzheng Liang were at Tencent Medical AI Lab for intern.

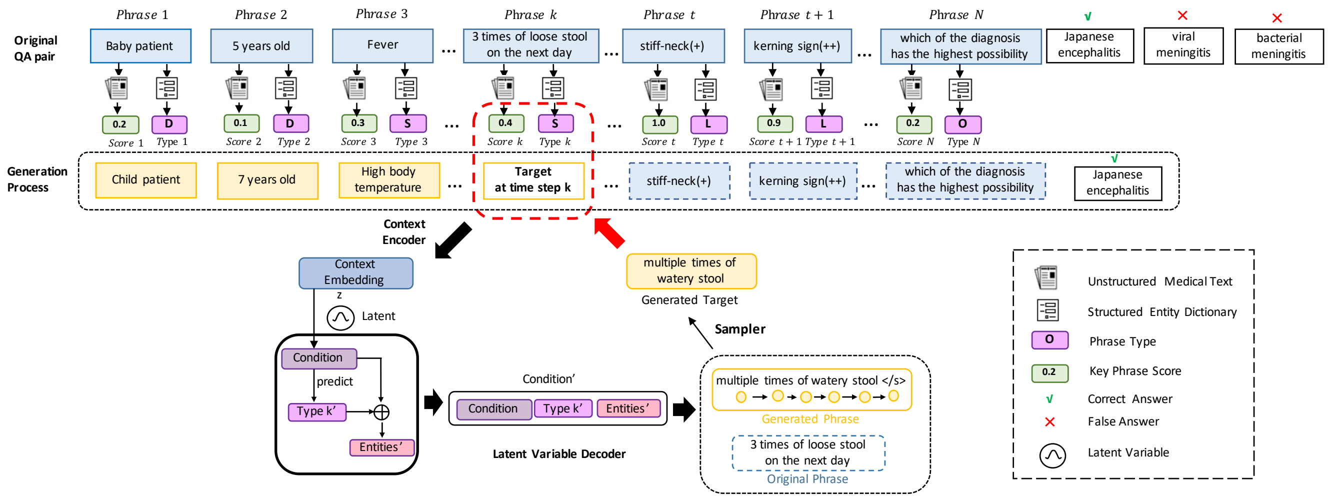

Figure 1: Overview of the proposed framework.

quantitative metrics, and the results confirm the high-quality of the generated medical QA pairs. Further, in an application of the proposed method to a medical certification exam, the experimental results show that the generated medical QA pairs improve the original QA system by six percent in terms of question-level accuracy.

Methodology

In this section, we introduce our framework for generating medical question-answer pairs based on existing pairs. Under the setting of medical QA, we assume that several questions can have the same answer, for example, the disease *Japanese encephalitis* can be matched with patients of *fever*, *pap test*(+), *stiff neck*(+) or *respiratory failure* due to the diversity of medical characteristics, while for a specific medical question, there is only one correct answer. Hence, we consider the generating process of medical QA pairs as generating questions for a certain answer. Technically speaking, our framework for generating medical QA pair can be considered as an approximation of the latent distribution of medical questions with the corresponding answer and sampling new questions from this latent distribution. As shown in Figure 1, the whole framework involves a key phrase detector and an entity-guided CVAE based Generator (eg-CVAE), which we will describe in details in the following subsections.

Key Phrase Detector

In order to approximate the unknown conditioned distribution of medical questions, we propose to exploit the intrinsic characteristic of medical questions that have the same answer with external knowledge. Specifically, every medical question Q consists of several phrases $P_k, k \in [1, N]$, such as patient's symptoms, examination results, etc. Each phrase is composed of several words. Among medical questions, there exist several phrases highly correlated with the answers, which offer us an insight that these phrases (denote as key phrases P'_k) should be kept with high probability in the generated new questions. Further, inspired by the entity type (Chang et al. 2018) in text generation, we assume that each phrase in the medical question has its own latent type information, and the structural relationships between phrases should also be kept in the generated questions.

As described above, key phrases of the question matter much to the corresponding answer in a medical QA pair. During the generation process, we should keep these key phrases with high probability. However, we do not know which phrases are key phrases for medical questions in prior. To solve this issue, we propose an unsupervised matching strategy to detect key phrase with the help of unstructured medical text. For each phrase P_k from the medical QA pair, we model the unsupervised key phrase detection problem as a matching problem. We assign each phrase with a normalized score $s_k, s_k \in [0, 1]$ which indicates the significance of the certain phrase towards the corresponding answer.

Rather than considering each phrase separately, we conjecture that the co-occurrence probability of a key phrase and the answer models the significance of that phrase. The co-occurrence information can be explored from the unstructured medical materials with the proposed method. Specifically, we perform an Elastic Search (Gormley and Tong) based retrieval over the medical materials for each medical QA pair, and the retrieved text is denoted as $T_i, i \in [1, N]$ (N stands for the number of retrieved text). Given that each T_i must contain the information of answer, the relevance of a certain phrase P_k with the corresponding answer can be calculated by matching the current phrase P_k with all the T_i.

Moreover, to mitigate the gap of deficient key phrases for training a matching model as (Wang, Hamza, and Florian 2017), we propose to utilize unsupervised word embedding with hierarchical pooling method. After learning word embedding from the massive unstructured medical text, each discrete word in both P and T can be represented with

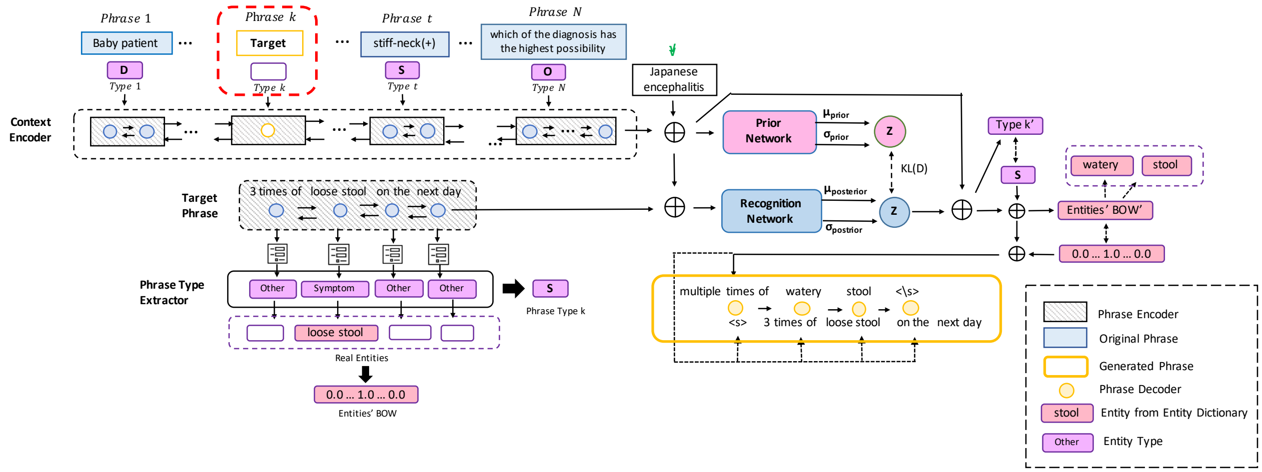

Figure 2: Entity-guided CVAE based Generator.

a real-valued vector v. Since each T_i can also be divided into splits as phrases P^{T_i} (each phrase consists of multiple words), we aim to represent each phrase with a fixed-length value vector and match them in the corresponding vector space. Inspired by (Henao et al. 2018), we perform a hierarchical pooling directly over the word embedding $v_j, j \in [1, L]$ in each phrase as follow: first, average pooling on $v_{j,j+k-1}, j \in [1, L-k+1]$ within each sliding window (size is k); then, max pooling on the induced average-pooling vectors. We match each phrase P_k in the original QA pair with the phrases P^{T_i} of T_i using cosine distance and store the highest score $s_k^{T_i}$. The unnormalized matching score for P_k with T is the mean value of $s_k^{T_i}, i \in [1, N]$. Then the scores for each phrase P_k in the QA pair will be normalized as the $s_k, s_k \in [0, 1]$ with the Min-Max method.

Entity-guided CVAE based Generator

A medical question has two levels of structures: one structure exists within a single phrase, which is dominated by local information of involved medical entities, and the other is a distinct across-phrase structure, which is characterized by aspects such as the corresponding answer and phrase types etc.. Thus rather than focusing on the word-level generation which is prevailing in dialog response generation and neural machine translation, we assume that the answer conditioned medical question generation can be modeled in a two-level hierarchy: sequences of sub-sequences (phrases), and sub-sequences of words.

Moreover, to explore the diversity of medical QA pair generation with constraints of external knowledge in the corresponding two levels, we introduce an idea based on Conditional Variational Autoencoder, which is constrained on the condition over the whole question and breaks the decoding process in each phrase with latent distribution z, Further inspired by human's process to generate a complete question (start from a scratch and then details), we model each phrase generation process as a three-pass decoding: first implicit type modeling, then explicit entities modeling, and finally phrase decoding. Figure 2 illustrates the procedure of the proposed Entity-guided CVAE based generator.

Conditional Variational Autoencoder . Enlightened by (Serban et al. 2017), we adapt the original CVAE for dialog generation to our setting by considering question generation as an iterative phrase generation process in Figure 2. To this end, we represent each phrase generation procedure with three random variables: the phrase context c, the target phrase x, and a latent variable z that is used to capture the latent distribution over all valid phrases. In our model, c is composed not only by the sequence of rest phrases but also the corresponding answer. We then define the conditional distribution $P(x, z|c) = P(x|c, z) \cdot P(z|c)$ and set the learning target is to approximate $P(z|c)$ and $P(x|c, z)$ via deep neural networks (parametrized by θ). We refer to $P_\theta(z|c)$ as the prior network and $P_\theta(x|c, z)$ as the target phrase decoder. Then the generative process of x is summarized as first sampling a latent variable z from $P_\theta(z|c)$ and then generating x by $P_\theta(x|c, z)$.

CVAE is trained to maximize the conditional log likelihood of x given c, meanwhile minimizing the KL divergence between the posterior distribution $P(x, z|c)$ and a prior distribution $P(z|c)$. We assume that z follows multivariate Gaussian distribution with a diagonal covariance matrix. Further, we introduce a recognition network $Q_\phi(z|x, c)$ to approximate the true posterior distribution $P(z|x, c)$. As proposed in (Sohn, Lee, and Yan 2015), CVAE can be efficiently trained with the *Stochastic Gradient Variational Bayes* (SGVB) framework (Kingma and Welling 2013) by maximizing the variational lower bound of the conditional log likelihood, which can be written as

$$\begin{aligned}\mathcal{L}(\theta, \phi; x, c) &= -KL(Q_\phi(z|x, c)||P_\theta(z|c)) \\ &\quad + E_{Q_\phi(z|x,c)}[\log P_\theta(x|c, z)] \leq \log P(x|c).\end{aligned} \tag{1}$$

At timestamp k of the whole QA generation process, the phrase encoder is a bidirectional recurrent neural network (Schuster and Paliwal 1997) with a gated recurrent unit

(GRU) to encode each phrase P_k into a fixed-size vector by concatenating the last hidden states of the forward and backward RNN $hv_k[\overrightarrow{hv_k}, \overleftarrow{hv_k}]$ as x. This basic phrase context encoder is a one-layer GRU network that encodes the $N-1$ context phrases (in training, the context phrases are from the original question; in testing, the preceding $k-1$ phrases are from the generated question) as $hv_{1:k-1}$ with $hv_{k+1:N}$. The last hidden state hv^c of the phrase context encoder is concatenated with the corresponding answer embedding a and $c = [hv^c, a]$. As we assume z follows isotropic Gaussian distribution, the recognition network $Q_\phi(z|x,c) \sim N(\mu, \sigma^2 I)$, the prior network $P_\theta(z|c) \sim N(\mu', \sigma'^2 I)$, and then we get:

$$\begin{bmatrix} \mu \\ \log(\sigma^2) \end{bmatrix} = W_r \begin{bmatrix} x \\ c \end{bmatrix} + b_r, \quad (2)$$

$$\begin{bmatrix} \mu' \\ \log(\sigma'^2) \end{bmatrix} = MLP_p(c). \quad (3)$$

The reparameterization trick (Kingma and Welling 2013) is adopted to obtain the samples of latent z from $N(z; \mu, \sigma^2 I)$ in training (recognition network) and from $N(z; \mu', \sigma'^2 I)$ in testing (prior network). The final phrase decoder at timestamp k is a one-layer GRU network with initial state set as $W_k[z, c] + b_k$. The words will be predicted sequentially by the phrase decoder.

Phrase-type Augmented Encoder Inspired by (Chang et al. 2018)'s insights to facilitate text generation by entity type information, we model the phrase type information of each phrase (insights employed by expert doctors such as lab examination and physical characteristics in real-world) in a medical question to dominate the structural characteristics of a medical question and further facilitate the generation.

However, it is difficult to acquire labeled data from professional doctors for phrase type information, and thus we propose to utilize the structured entity dictionary to extract such information. Specifically, we assume that each phrase information to be influenced by two levels of characteristics: 1) global characteristic as the surrounding or context phrases' type information; 2) local characteristic as entity type information within each phrase. The first characteristic indicates as (Peters et al. 2018) proposed in word embedding, and the type information can be contextualized rather than static.

To this end, we design a sequence labeling task for pretraining, whose learning goal is to predict each words' type (for those words not in the entity dictionary, the type is considered as "other") over the whole question precisely. Thus the pre-training task can incorporate both local entities' type information and global context type information. A Bi-LSTM-CRF model, which takes each words' embedding in the question as input and their types as output, is applied to the pre-training task. We use Bi-LSTM layer to encode word-level local features, and CRF layer to capture sentence-level type information. As the pre-training task's accuracy can achieve 97.08%, we assume that the hidden states of Bi-LSTM for each word k as $h_k[\overrightarrow{h_k}, \overleftarrow{h_k}]$ can encode the contextualized type information for each word. Considering that each phrase can be split into multiple words, the phrase type information is introduced by performing max-pooling over each words' h_k.

We further concatenate contextualized type vector t_k at timestamp k for generating phrase P_k to hv_k as $hv'_k = [hv_k, t_k]$. t_k is pre-trained with the previous sequence labeling task and different at each timestamp of the whole medical QA generation procedure. In this way, the new $x' = hv'_k$ will be applied for the recognition network.

Entity-guided Decoder Other than only conditioning on the corresponding answer, we introduce an extra constraint on latent z to keep it meaningful during generating process, in order to incorporate the structural feature of the question itself, namely the type information introduced by entity dictionary, and to ensure that generated questions match well with answers. This can also guarantee that the relationship between phrases-type information will be kept to the most extent during the procedure of generating new phrases.

(Xia et al. 2017) points out that when people perform the task of generating questions/writing texts, they tend to start from scratch and add details step by step. This can be considered as a multiple pass decoding procedure. For our scenario, modeling entity explicitly can contribute to the training of decoder, which is coherent with (Chang et al. 2018)'s work on generating recipes and codes. At the phrase level, we conjecture that entities can be regarded as the skeleton, and there is no fixed linguistic sequence of these entities in a phrase. Thus, to incorporate this multiple pass generating procedure, we propose a model that first decodes entities e in a phrase. Since we add variations at entity-level, via producing different sentences with diverse entities that have the same meaning, diversities are ensured in our final generation process.

To this end, we assume that the generation of phrase P_k as x depends on c, z, t and e; e relies on c, z, t; and t relies on c, z. To be more specific, during training, the initial state of the final decoder is $s_k = W_k[z, c, t, e] + b_k$ and the input is $[w_{1:n^k}, t, e_k]$ where $w_{1:n^k}$ is the word embedding of words in x and e_k is average pooling embedding of the entire entity embedding in x. In the one-pass type modeling, there is an MLP to predict $t' = MLP_t(z, c)$ based on z and c. In the two-pass entity modeling, another MLP is first introduced to predict $e_{\text{softmax}'} = MLP_e(z, c, t)$ based on z, c and t. Then $e_{\text{softmax}'}$ is multiplied with the whole entity embedding matrix for the aggregation of the e'_k. In the testing stage, the predicted t' and e'_k is used by the final phrase decoder.

Training Objective

In order to induce meaningful latent variable z, we explicitly model the generation of x as a multi-pass process, which can relieve the vanishing latent variable problem in the straightforward training of VAE with RNN decoder.

Specifically, by introducing phrase-type information during the first pass of decoding from scratch to details, we suppose that the generation of x is based on c, z and t; and t is based on c. Then the modified variational lower bound for eg-CVAE is as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi;x,c,t) &= -KL(Q_\phi(z|x,c,t)||P_\theta(z|c)) \\ &+ E_{Q_\phi(z|x,c,t)}[\log P_\theta(t|c,z)] \\ &+ E_{Q_\phi(z|x,c,t)}[\log P_\theta(x|c,z,t)].\end{aligned} \tag{4}$$

To refine phrase-type information into detailed entities in the second pass, we model e explicitly based on the assumption that the produce of x is divided into two phases: exploiting phrase-type to generate e; and using e, t, c and z to generate x. Thus the final eg-CVAE model is trained by maximizing:

$$\begin{aligned}\mathcal{L}(\theta,\phi;x,c,t,e) &= -KL(Q_\phi(z|x,c,t,e)||P_\theta(z|c)) \\ &+ E_{Q_\phi(z|x,c,t,e)}[\log P_\theta(t|c,z)] \\ &+ E_{Q_\phi(z|x,c,t,e)}[\log P_\theta(e|c,z,t)] \\ &+ E_{Q_\phi(z|x,c,t,e)}[\log P_\theta(x|c,z,t,e)].\end{aligned} \tag{5}$$

Furthermore, the KL annealing (Serban et al. 2016a) technique as gradually increasing the weight of the KL term from 0 to 1 during training and auxiliary bag-of-words loss of x proposed by (Zhao, Zhao, and Eskenazi 2017) are also adopted.

Experiments

Dataset

To validate the effectiveness of the proposed method, we collect real-world medical QA pairs from the National Medical Licensing Examination of China (denoted as NMLEC_QA). The collected NMLEC_QA dataset contains $18,798$ QA pairs, and we aim to generate new QA pairs based on these original ones. The medical entity dictionary is extracted from medical Wikipedia-style pages[1], and the constructed dictionary covers 19 types of medical entities. The unstructured medical materials consists of $2,130,128$ published paper in medical domain and 518 professional medical textbooks.

Implementation Details

The proposed method is trained with the following hyperparameters: Word embedding (Mikolov et al. 2013) is pretrained using the whole unstructured medical materials with a vector dimension of 200, and the learned vector representations are shared across different components of the proposed method. The phrase encoder's hidden dimension is set to be 300. The hierarchical context phrase encoder has a hidden dimension of 600, and the latent variable z has a size of 200. The number of retrieved medical text is set to be 10. The size of sliding window in hierarchical pooling method is set to 3. Both the prior network and the MLP for one-pass type decoder have one hidden layer of size 400 and tanh nonlinearity activation function. The two-pass entity decoder is another MLP with the dimension of the entity vocabulary size. By connecting to a softmax layer, an entity embedding with a dimension of 50 is then applied for aggregation. The final phrase decoder's hidden dimension is set to be 400. The initial weights of these networks are sampled from a uniform distribution $[-0.08, 0.08]$. The models are trained with a mini-batch size of 30, Adam optimizer with a learning rate of 0.001, and gradient clipping at 5. Further, we use the BOW loss along with KL annealing of $10,000$ batches. We conduct these parameter selections based on the variational lower bound.

[1] `http://www.xywy.com/`

Baselines

We compare the performance of the proposed method **eg-CVAE** with two recently-proposed text generation methods: *HRED* (Serban et al. 2016b), a generalized sequence-to-sequence model with hierarchical RNN encoder, and *CVAE* (Serban et al. 2017), a hierarchical conditional VAE model with KL-annealing. We also test the proposed method in the ablation of the type modeling or the entity modeling: *type-CVAE* with type decoding as the first-pass, and *entity-CVAE* with entity decoding as the first-pass.

Evaluation based on Automatic Metric

Automatically evaluating the quality of generated text remains challenging (Liu et al. 2016), and thus we design automatic evaluation metrics by considering our specific scenario. As mentioned above, we assume that each QA pair can be considered as a question sampled from a latent distribution conditioned on the corresponding answer. To be more specific, based on each original question-answer pair, we generate N new questions by iteratively sampling candidate phrases and choosing phrases in a beam-search way (Sutskever, Vinyals, and Le 2014). As the generation procedure is at phrase-level, we evaluate each generated question by comparing the generated phrases with the corresponding original phrases and averaging the evaluation results over all the phrases in the questions.

We adopt the following three standard metrics to measure the quality of the generated questions from lexical, semantic and diversity perspective.

- *Smoothed Sentence-level BLEU* (Chen and Cherry 2014): this is a lexical similarity metric that measures the degree to which a generated phrase contains n-gram overlaps with the original phrase with a length penalty. As N new questions are generated, we define the n-gram precision and n-gram recall as the average and the maximum value of N n-gram BLEU scores respectively. We use 3-gram with smoothing technique, and BLEU scores are normalized to $[0, 1]$.
- *Cosine similarity of Bag-of-words (BOW) embeddings*: this is a metric to match phrase embeddings of two phrases through the average, extrema or greedy strategy over all the word embeddings as follows respectively: *Average*: cosine similarity between the averaged word embeddings; *Extrema* (Forgues et al. 2014): cosine similarity between the biggest extreme values among the word embeddings of the two phrases; *Greedy* (Rus and Lintean 2012): matching words in two phrases greedily based on their embeddings' cosine similarity and averaging the obtained scores.

Table 1: Performance comparison under automatic evaluation metrics.

Method	BLEU			BOW Embedding			intra-dist		inter-dist	
	Precision	Recall	F1	Average	Extrema	Greedy	dist-1	dist-2	dist-1	dist-2
HRED	0.435	0.737	0.547	0.753	0.705	0.809	0.837	0.912	0.205	0.255
CVAE	0.454	0.705	0.533	0.863	0.872	0.887	0.803	0.991	0.562	0.538
type-CVAE	0.507	0.748	0.572	0.872	0.852	0.892	0.831	**0.997**	0.555	0.581
entity-CVAE	**0.541**	**0.781**	**0.613**	**0.891**	**0.903**	**0.874**	0.840	0.996	0.533	0.554
eg-CVAE	0.450	0.611	0.494	0.802	0.793	0.819	**0.867**	0.994	**0.637**	**0.589**

- *Distinct* (Gu et al. 2018): this is an approach to compute the diversity of the generated phrases. The ratio of unique n-grams over all n-grams in the generated phrases is denoted as distinct-n. We further define *intra-dist* as the average of distinct values within each sampled phrase and *inter-dist* as the distinct value among all sampled phrases.

We compare the proposed method eg-CVAE with the aforementioned baselines on the collected real-world dataset NMLEC_QA, and report the results based on the automatic evaluation metrics in Table 1. The highest score in each column is in bold for clarity. In the following, we discuss the experiment results in details.

First, we examine the results from the perspective of semantic similarity in terms of BLEU score and BOW score. When we design the proposed method eg-CVAE, diversity is promoted, and thus its semantic similarity is not high. The vanilla CVAE does not involve any constraint on the latent distribution of z, and the HRED (Serban et al. 2016b) models the decoding process in a definite way without further manipulation on the hidden context, so their semantic similarity scores are medium. A variant of the proposed method type-CVAE models type information explicitly, and another variant entity-CVAE models entity explicitly. These facilitate models to generate QA pairs that are similar to the original.

On the other hand, from the perspective of diversity, the proposed method eg-CVAE has the highest score in terms of distinct metrics. This is because that we hierarchically generate new questions based on the latent answer-conditioned distribution, rather than a definite decoding process. As pointed out in (Serban et al. 2017), this hierarchical strategy can prevent diversity being injected at the low level.

As we discussed in the Introduction section, the method to generate medical QA pairs should have both diversity and validity, and we show the effectiveness of the proposed method eg-CVAE here. Besides these quantitative evaluation results, we also conduct extensive case studies, which also confirm the advantage of the proposed method. Due to the page limitation, case studies are included in the supplementary file.

Evaluation based on Human Annotated Metric

The above automatic evaluation metrics focus on the linguistic aspects of the generated QA pairs. Here we further conduct experiments to ensure that the generated medical questions are matched with the corresponding answers, with the help of human annotation.

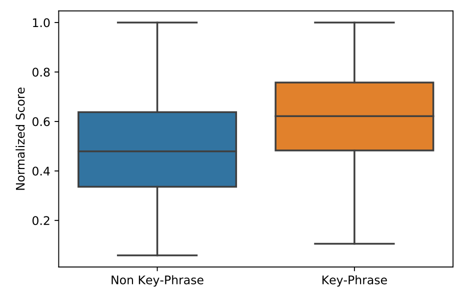

Figure 3: The distribution of proposed score for key phrases.

Traditional human evaluation requires both expensive and intensive work for experts to directly verify whether each generated medical QA pairs matches. Also, medical QA pairs produced by each method needs to be relabeled, which is time-consuming and might not be feasible. To solve these challenges, we propose a reusable solution for human annotation.

As mentioned above, there exist some key phrases that are highly related to the answer in each medical question. So we assume that the probability of whether these key phrases are kept in the generated question represents the probability of whether the generated question matches the corresponding answer. We randomly sampled 500 medical QA pairs from NMLEC_QA dataset, and then employ three experts (doctors) to label the key phrases in these questions. The labeled key phrases with at least two experts' consensus will have the final label of "Yes", others "No".

In the proposed method eg-CVAE, we have a component for key phrase detection, which only utilizes unstructured medical texts without any supervision. It assigns each phrase a score (a probability) to be a key phrase. The distribution of the key phrase score of labeled data is plotted in Figure 3. From this figure, we can observe that the unsupervised key phrase detector could assign the real key phrases with higher scores, which can ensure those key phrases unchanged in the generated questions with high probability. In this way, the proposed method can generate medical questions that match the conditioned answers.

Table 2: Usefulness of the generated QA pairs.

Dataset	Accuracy
Original	61.97
+ HRED QA	58.78
+ CVAE QA	62.28
+ type-CVAE QA	65.27
+ entity-CVAE QA	64.67
+ eg-CVAE QA	**67.96**

Evaluation on a QA System

To further study the usefulness of the generated medical QA pairs, we integrate such generated pairs into a QA system, which is an attention-based model (Cui et al. 2017) for NMLEC_QA dataset. The results are summarized in Table 2.

First of all, integrating the generated QA pairs from HRED method leads to the lowest accuracy, which is even lower than the one only based on the original dataset. As pointed out in (Serban et al. 2017), HRED is very likely to favor short-term predictions instead of long-term predictions. Rather than globally considering context phrases to generate a meaningful phrase for the current slot, the model tends to repeat the predicted correct word. The lack of diversity and repeat of common words lead to the difference in the generated questions' distribution and the original one, which may hurt the performance of the QA system and introduce noise to the original dataset.

The generated QA pairs from the vanilla CVAE only help the original QA system to achieve a limited improvement. This is caused by the fact that there is no constraint on the latent distribution, and the guidance from the corresponding answer and question structure becomes weak during the whole generation process.

Two variants of the proposed method, entity-CVAE and type-CVAE, generated QA pairs that seem to boost the original QA system with significant improvement. Each of them introduces external constraints on the latent variable in the decoding phase, which may help to diversify the generated questions while keeping linguistic and structural relationships within original questions. Furthermore, type-CVAE generates QA pairs that seem to be more helpful to the QA system. This benefit may come from the modeling of type information, which allows the generated questions to be relatively more diverse and thus introduces more useful knowledge. The proposed method eg-CVAE combines the advantages of entity-CVAE and type-CVAE, building a three-pass decoding process, and thus improves the QA system to achieve the highest accuracy. These observations demonstrate the usefulness of the generated medical QA pairs by the proposed method.

Related Work

Question Generation (Rus et al. 2010) has attracted increasing attention in recent years. However, many existing work only focuses on the similarity of generated questions with the original ones, but ignores the usefulness of generated questions with the corresponding answers. Earlier work in question generation employed rule-based approaches to transform input texts into corresponding questions, usually requiring some well-designed general rules (Mitkov and Ha 2003), templates (Labutov, Basu, and Vanderwende 2015; Chali and Golestanirad 2016) or syntactic transformation heuristics (Agarwal and Mannem 2011; Ali, Chali, and Hasan 2010). Recent studies leveraged neural networks to generate questions in an end-to-end fashion. (Du, Shao, and Cardie 2017) applied the attention-based sequence-to-sequence model to generate questions in the context of reading comprehension.

Other existing work, which tackles the usefulness and models the question-answer pair generation directly, still sets the diversity of questions for the corresponding answer aside and requires related context in prior. (Serban et al. 2016a) applied the encoder-decoder framework to generate question-answering pairs from built knowledge base triples. (Subramanian et al. 2018) formulated the question-answer pair generation in reading comprehension, where each pair is provided with a given high-quality context and the answer is a text span of the context, separately with the answer detection and question generation problem. Coreference knowledge is also introduced for question-answer pair generation with Wikipedia articles as the context in (Cardie and Du 2018). (Yuan et al. 2017) leveraged policy gradient techniques to further improve the generation quality in reading comprehension. (Duan et al. 2017) investigated to integrate generated questions from given context to the question-answering system on sentence selection task, which leveraged both rule-based features and neural network to approximate the semantics of generated questions with original ones.

Compared to existing work, our work tackles a specific scenario of QA pairs generation for the medical domain, which does not involve any prior context. To ensure the validity of generated QA pairs, we proposed an unsupervised detector to automatically explore external materials. We also proposed to model the question-answer pair generation problem directly as approximating the latent distribution of medical questions with the corresponding answer.

Conclusions

In this paper, we introduced a novel framework, consisting of an unsupervised key phrase detector and an Entity-guided CVAE-based generator, for automated question-answer pair generation in the medical domain. Different from existing seq2seq models that involve definite encoding-decoding procedure to restrict the generation scope, or traditional CVAE models that directly approximate the posterior distribution over the latent variables to a simple prior, the proposed method models the generation process as a multi-pass procedure (type, entity and phrase as constraints over the latent distribution) to ensure both validity and diversity. Experiments on a real-world dataset from the National Medical Licensing Examination of China demonstrate that the proposed method outperforms existing methods and can generate more diverse, informative and valid medical QA pairs that further benefit the examination QA system.

References

[Agarwal and Mannem 2011] Agarwal, M., and Mannem, P. 2011. Automatic gap-fill question generation from text books. In *Proceedings of EMNLP Workshop*, 56–64.

[Ali, Chali, and Hasan 2010] Ali, H.; Chali, Y.; and Hasan, S. A. 2010. Automation of question generation from sentences. In *Proceedings of QG Workshop*, 58–67.

[Cardie and Du 2018] Cardie, C., and Du, X. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *Proceedings of ACL*, 1907–1917.

[Chali and Golestanirad 2016] Chali, Y., and Golestanirad, S. 2016. Ranking automatically generated questions using common human queries. In *Proceedings of INLG*, 217–221.

[Chang et al. 2018] Chang, K.; Parvez, M. R.; Chakraborty, S.; and Ray, B. 2018. Building language models for text with named entities. In *Proceedings of ACL*, 2373–2383.

[Chen and Cherry 2014] Chen, B., and Cherry, C. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of ACL Workshop*, 362–367.

[Cui et al. 2017] Cui, Y.; Chen, Z.; Wei, S.; Wang, S.; Liu, T.; and Hu, G. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the ACL*, 593–602.

[Du, Shao, and Cardie 2017] Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: neural question generation for reading comprehension. In *Proceedings of ACL*, 1342–1352.

[Duan et al. 2017] Duan, N.; Tang, D.; Chen, P.; and Zhou, M. 2017. Question generation for question answering. In *Proceedings of EMNLP*, 866–874.

[Forgues et al. 2014] Forgues, G.; Pineau, J.; Larchevêque, J.-M.; and Tremblay, R. 2014. Bootstrapping dialog systems with word embeddings. In *Proceedings of NIPS Workshop*.

[Gormley and Tong] Gormley, C., and Tong, Z. *Elasticsearch: The definitive guide: A distributed real-Time search and analytics engine*.

[Gu et al. 2018] Gu, X.; Cho, K.; Ha, J.; and Kim, S. 2018. Dialogwae: multimodal response generation with conditional wasserstein auto-encoder. *CoRR* abs/1805.12352.

[Henao et al. 2018] Henao, R.; Li, C.; Carin, L.; Su, Q.; Shen, D.; Wang, G.; Wang, W.; Min, M. R.; and Zhang, Y. 2018. Baseline needs more love: on simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of ACL*, 440–450.

[Kingma and Welling 2013] Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Labutov, Basu, and Vanderwende 2015] Labutov, I.; Basu, S.; and Vanderwende, L. 2015. Deep questions without deep understanding. In *Proceedings of ACL*, 889–898.

[Liu et al. 2016] Liu, C.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of EMNLP*, 2122–2132.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of ACL*, 3111–3119.

[Mitkov and Ha 2003] Mitkov, R., and Ha, L. A. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of NAACL-HLT Workshop*, 17–22.

[Peters et al. 2018] Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, 2227–2237.

[Rus and Lintean 2012] Rus, V., and Lintean, M. C. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of NAACL-HLT Workshop*, 157–162.

[Rus et al. 2010] Rus, V.; Wyse, B.; Piwek, P.; Lintean, M. C.; Stoyanchev, S.; and Moldovan, C. 2010. The first question generation shared task evaluation challenge. In *Proceedings of INLG*.

[Schuster and Paliwal 1997] Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 2673–2681.

[Serban et al. 2016a] Serban, I. V.; García-Durán, A.; Gulcehre, C.; Ahn, S.; Chandar, S.; Courville, A.; and Bengio, Y. 2016a. Generating factoid questions with recurrent neural networks: the 30m factoid question-answer corpus. In *Proceedings of ACL*, 588–598.

[Serban et al. 2016b] Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*, 3776–3784.

[Serban et al. 2017] Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of AAAI*, 3295–3301.

[Sohn, Lee, and Yan 2015] Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In *Proceedings of NIPS*, 3483–3491.

[Subramanian et al. 2018] Subramanian, S.; Wang, T.; Yuan, X.; Zhang, S.; Trischler, A.; and Bengio, Y. 2018. Neural models for key phrase extraction and question generation. In *Proceedings of ACL Workshop*, 78–88.

[Sutskever, Vinyals, and Le 2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, 3104–3112.

[Wang, Hamza, and Florian 2017] Wang, Z.; Hamza, W.; and Florian, R. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of IJCAI*, 4144–4150.

[Xia et al. 2017] Xia, Y.; Tian, F.; Wu, L.; Lin, J.; Qin, T.; Yu, N.; and Liu, T.-Y. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Proceedings of NIPS*, 1784–1794.

[Yuan et al. 2017] Yuan, X.; Wang, T.; Gülçehre, Ç.; Sordoni, A.; Bachman, P.; Zhang, S.; Subramanian, S.; and Trischler, A. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of ACL Workshop*, 15–25.

[Zhao, Zhao, and Eskenazi 2017] Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of ACL*, 654–664.

Supplementary Material: Case Study

To qualitative analyze the proposed method through some real cases, we first compare the generated QA pairs from different models in Figure 4. Secondly, we present a detailed illustration of our proposed method entity-guided CVAE's generation process in Figure 5.

Case study for generated QA pairs from different models

We demonstrate several generated medical QA pairs from different models in Figure 4. Each example is consisted of an original valid QA pair and three generated questions, which are sampled based on the raw one through beam-search. Three models are compared with including HRED, CVAE and eg-CVAE.

1. For HRED, it is easy to find that the generated questions' diversity is limited since the model tends to repeat the seed phrases (e.g., the meaningless repetition of "RBC" and "anxiety") and the important information describing topographic shape (e.g., "lower than" in "HB is lower than normal") is easily lost. Since the model does not distinguish the phrases between each other, the lexical metric (BLEU score) which measures n-gram exact match with the original question seems to be high, even though the generated question and the corresponding answer do not match.

2. For CVAE, it is obvious that CVAE explores the diversity in generation. However, the sampled questions show that ambiguous phrases are often generated in a key place. For instance, in the first sentence "wbc 3.45×10^{12}/l" is very likely to indicate inflammation, in the second sentence excessive symptoms of "diarrhea" may guide to either anemia or diarrhea, and in the third sentence "sudden fever after menstruation, discomfort" in most cases indicates endocrine disorders rather than anemia. This is due to two reasons - the model involves limited constraints on the latent distribution from the answer and the iterative generation setting makes the model focus more on the previous generated phrases, which to some extent weakens the restriction from answer on the whole generation.

3. For eg-CVAE, we can clearly see that this model retains the one-to-many diversity property of each phrase's generation. Moreover considering the validity, the generated imperative semantics of the non-key phrases are consistent with the implicit semantics of the original questions of anemia. For example, although the semantics of "the poor face", "anxious" and "whitish complexion" are different, they does not influence on the overall diagnosis of "anemia". For more examples, "the normal systolic blood pressure" and "normal liver" have no influence on the judgment of "anemia" as they are both normal body sign.

Detailed case study for generation process of entity-guided CVAE

To further demonstrate the advantages of the proposed eg-CVAE in terms of diversity and validity in the above analysis, we present in detail the generation process involving different constraints with respect to the latent variable z (multiple-pass decoding procedure) in Figure 5.

From the results, on the one hand, we can observe that explicit entity modeling at first-pass makes the generated phrases strongly related to the modeled entities. Many en-CVAE generated phrases directly contain the modeled entities, and the diversity is relatively limited. Moreover, once the decoded entities are relatively abstractive, (e.g., "poor body"), the generated phrase may not contain the key information in the original question, such as informative phrase "HB is lower than normal" replaced by trivial phrase "low spirit". On the other hand, implicit type modeling at first-pass encourages more diversity in generation. Since the constraint extent on type by decoding the contextualized type vector is much looser than model with the decoding explicit entities, the generated diversity will be more broad, such as "occasional ventosity" or "yellow complexion", etc.

To handle this phenomenon, eg-CVAE comprehensively treats explicit entity modeling and implicit type modeling as different decoding passes. By first modeling type information and then introducing it as a priori to the entity modeling, eg-CVAE prevents the loss of key information; and by adding variants through multiple-pass decoding processes, generated questions are well diversified. In this way, the diversity and validity of generated QA pairs are both guaranteed.

Original QA	**Q**	男,1岁,脸色苍白1个月,易疲倦,时而焦虑,胃纳差,体检:肝脏位于肋下3cm,质中,血红蛋白(HB)低于正常, **红细胞数(RBC)低于正常,平均红细胞体积(MCV)低于正常, mch 20pg, mchc 0.26**,最可能的诊断是 Male, 1 year-old, face been pale for 1 month, often tired, sometimes anxious, poor appetite, examination result: liver is 3 cm below the ribs, between hard and soft, hb is lower than normal, **rbc is lower than normal. mcv is lower than normal, mch 20pg, mchc 0.26**, the most likely diagnosis is
	A	贫血 (anemia)
HRED	**Q1**	女,1岁,脸色苍白1个月,焦虑,焦虑,体检,肝脏肋下未触及,HB正常, RBC RBC 1.02 正常, MCV正常,mch 10pg,mchc 10.3cm,最可能的诊断是 Female, 1 year-old. face been pale for 1 month, anxiety, anxiety, examination result: untouched liver under the ribs, hb normal, rbc count rbc count 1.02 normal, mcv normal, mch 10pg, mchc 10.3cm, the most likely diagnosis is
	Q2	男,1岁,脸色 3个月,焦虑,焦虑,焦虑,体检,肝脏肋下,HB 正常, RBC 1.02 正常, MCV 正常,mch 0.1pg,mchc 0.26,最可能的诊断是 Male, 1 year-old, face 3 months, anxiety, anxiety, anxiety, examination result: liver under the ribs, hb normal, rbc count 1.02 normal, mcv normal, mch 0.1pg, mchc 0.26, The most likely diagnosis is
	Q3	女,1岁,脸色苍白1个月,焦虑,焦虑,体检,肝脏肋下未,体温39.3oC,HB高于正常, RBC 1.02正常, MCV正常, mch 10pg, mchc 0.26,最可能的诊断是 Female, 1 year-old, face has been pale for 1 month, anxiety, anxiety, examination result: liver under the ribs, body temperature 39.3oC, hb higher than normal, rbc count 1.02 normal, mcv normal, mch 10pg, mchc 0.26, the most likely diagnosis is
CVAE	**Q1**	男,58岁,头发少1个月,易疲倦,全身无力,体检脸色发黄,肝脏肋下2.5cm,胃纳差,尿蛋白(+),wbc 3.45×10^{12}/l, MCV低于正常, mch 24 pg, mchc 0.26, 最可能的诊断是 Male, 58 year-old, hair deficiency for 1 month, easy to get tired, general weakness, examination result: yellow complexion, liver is 2.5cm under the ribs, poor storage of stomach, urine protein (+), wbc 3.45x10^{12}/l, mcv lower than normal, mch 24pg, mchc 0.26, the most likely diagnosis is
	Q2	女,4岁,发热,伴面色污秽,腹泻腹痛1天,查体,双腿呈非凹陷水肿,大便WBC高于正常,尿中WBC高于正常,MCV低于正常, mch 20pg, mchc 0.26,最可能的诊断是 Female, 4 year-old, fever, with facial stains, diarrhea, abdominal pain for 1 day, examination result: legs with non-sink edema, wbc in stool is higher than normal, the number of wbc in unary is higher than normal, mcv below normal, mch 20pg, mchc 0.26, the most likely diagnosis is
	Q3	女,21岁,脸色苍白1个月,经后4天,偶有轻微不适,月经后突发热,不适,偶有下腹痛,质硬,体检脸色发黄,HB低于正常,RBC低于正常,MCV低于正常, mch 20pg, mchc 0.26,最可能的诊断是 Female, 21 year-old, face been pale for 1 month, 4 days after menstruation, mild discomfort occasionally, sudden fever after menstruation, discomfort, occasional lower abdominal pain, hard, examination result: yellow complexion, hb below normal, rbc count lower than normal, mcv lower than normal, mch 20pg, mchc 0.26, the most likely diagnosis is
Eg-CVAE	**Q1**	患儿,7岁,脸色苍白,时有吞咽不下,精神萎靡伴焦虑,一周加重,查体,身高83cm,肝脏外观质硬,HB低,RBC低于正常,MCV低于正常,mch 20pg,mchc 0.26,最可能的诊断是 Child patient, 7 year-old, pale complexion, unable to swallow occasionally, mental wilting with anxiety, a week of aggravation, examination result: height is 83cm, liver hard, low hb, rbc count is below normal, mcv is lower than normal, mch 20pg, mchc 0.26, the most likely diagnosis is
	Q2	患者,男,10岁,脸色发黄1个月,时有腹泻,伴焦虑,有时加重,体检,肝脏位于肋下2cm,周身帮增生,HB低于正常,RBC低于正常,MCV低于正常,mch 20pg,mchc 0.26,最可能的诊断是 Patient, male, 10 year-old, Face has been yellow for 1 month, diarrhea occasionally, with anxiety, sometimes aggravates, examination result: the liver is 2cm below the ribs, surrounding tissue proliferate , the hb is lower than normal, rbc count lower than normal, mcv lower than normal, mch 20pg, mchc 0.26, the most likely diagnosis is
	Q3	男,8岁,脸色差,时有疲倦、焦虑,查体,收缩压正常、舒张压正常,脾脏正常,外观毛玻璃样,HB60k/L,rbc 3.3×10^{12}/l, MCV低于正常, mch 20pg, mchc 0.26,最可能的诊断是 Male, 8 year-old, with poor face, feel tired occasionally, anxious, examination result: normal systolic blood pressure, normal diastolic blood pressure, normal spleen, appearance of ground glass, hb 60k/L, rbc 3.3x10^{12}/l, mcv is lower than normal, mch2 0pg, mchc 0.26, the most likely diagnosis is

Figure 4: Case study for generated QA pairs of different methods (the key phrases in original QA pair are in bold)

Original	**Q**	男 Male	1岁 1 year-old	脸色苍白1个月 face been pale for 1 month	易疲倦 easy to get tired	胃纳差 poor appetite	体检: examination result	肝脏位于肋下3cm liver is 3cm under the rib	质中 between soft and hard	HB (<)	**RBC (<)**	**MCV (<)**	**Mch 20pg**	**Mchc 0.26**	最可能的诊断是 the most possible diagnosis is
	L	0	0	0	0	0	0	0	0	0	**1**	**1**	**1**	**1**	0
	S	0.438	0.407	0.631	0.157	0.411	0.223	0.207	0.615	0.836	0.748	0.833	0.770	1.000	0.141
En-CVAE	**Q**	患儿 Child patient	1岁 1 year-old	脸色苍白1个月 face been pale for 1 month	时而呈兴奋、全身无力 sometimes excited, general weakness	时而焦虑 occasional anxiety	时而个月来面色部剧烈疼痛 face been hurting strongly for 1 month	肝脏略于正常 liver kind of normal	略淡染区 light infection area	精神萎靡 low spirit	RBC(<)	MCV (<)	Mch 20pg	Mchc 0.26	最可能的诊断是 the most possible diagnosis is
	E	患儿 child patient	岁 year-old	脸色 face	全身无力 general weakness	疼痛 pain	正常 normal	正常 normal	正常 normal	体质差 poor body		-	-	-	-
Tp-CVAE	**Q**	女 Female	1岁多 more than 1 year-old	脸色发黄1个月 yellow complexion for 1 month	精神萎靡1个月 spirit been low for 1 month	时而出现肚子胀 occasional ventosity	体检发现 examination result found	肝脏位正常 normal liver location	质软 soft	HB(<)	RBC 2.45×10^{12}	MCV (<)	Mch 26pg	Mchc 0.26	最可能的诊断是 the most possible diagnosis is
	T	T1	T1	T6	T1	T3	T3	T6	T2	T2	T2	T4	T4	T4	T1
Eg-CVAE	**Q**	男 Male	8岁 8 year-old	脸色差 poor complexion	精神萎靡 low spirit	焦虑 anxiety	查体 physical examination	肝脏位正常 normal liver location	外观毛玻璃样 appearance of ground glass	HB 60k/L	RBC 3.3×10^{12}	MCV (<)	Mch 26pg	Mchc 0.26	最可能的诊断是 the most possible diagnosis is
	T	T1	T1	T6	T1	T3	T3	T6	T2	T2	T2	T4	T4	T4	T1
	S	男 male	岁 year-old	脸色 face	精神萎靡 low spirit	焦虑 anxiety	检查 exam	正常 normal	肝脏 liver	血红蛋白 hb	红细胞 rbc	-	-	-	-

Figure 5: Further case study for generated QA pairs of eg-CVAE. (Q, L, S, T and E stand for question, label, score, type and entity, respectively. (<) indicates that "lower than normal")