

# Generative Question Refinement with Deep Reinforcement Learning in Retrieval-based QA System

Ye Liu

Department of Computer Science,  
University of Illinois at Chicago, USA  
yliu279@uic.edu

Chenwei Zhang

Amazon, Seattle, WA, USA  
cwzhang@amazon.com

Xiaohui Yan

Poisson Lab, Huawei Technologies,  
Shenzhen, China  
yanxiaohui2@huawei.com

Yi Chang

School of Artificial Intelligence, Jilin  
University, Jilin, China  
yichang@jlu.edu.cn

Philip S. Yu

Department of Computer Science,  
University of Illinois at Chicago, USA  
psyu@uic.edu

## ABSTRACT

In real-world question-answering (QA) systems, ill-formed questions, such as wrong words, ill word order and noisy expressions, are common and may prevent the QA systems from understanding and answering accurately. In order to eliminate the effect of ill-formed questions, we approach the question refinement task and propose a unified model, QREFINE, to refine the ill-formed questions to well-formed questions. The basic idea is to learn a Seq2Seq model to generate a new question from the original one. To improve the quality and retrieval performance of the generated questions, we make two major improvements: 1) To better encode the semantics of ill-formed questions, we enrich the representation of questions with character embedding and the contextual word embedding such as BERT, besides the traditional context-free word embeddings; 2) To make it capable to generate desired questions, we train the model with deep reinforcement learning techniques that consider an appropriate wording of the generation as an immediate reward and the correlation between generated question and answer as time-delayed long-term rewards. Experimental results on real-world datasets show that the proposed QREFINE can generate refined questions with high readability but fewer mistakes than original questions provided by users. Moreover, the refined questions also significantly improve the accuracy of answer retrieval.

## CCS CONCEPTS

• **Information systems** Information retrieval query processing; Question answering; Information retrieval; Query representation.

Work done while the second author at University of Illinois at Chicago.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '19, November 3–7, 2019, Beijing, China  
© 2019 Association for Computing Machinery.  
ACM ISBN 978-1-4503-6976-3/19/11...\$15.00  
<https://doi.org/10.1145/3357384.3358046>

## KEYWORDS

Question Refinement; Off-policy Reinforcement Learning; Multi-grain Word Embedding

### ACM Reference Format:

Ye Liu, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2019. Generative Question Refinement with Deep Reinforcement Learning in Retrieval-based QA System. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*, November 3–7, 2019, Beijing, China. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3357384.3358046>

## 1 INTRODUCTION

QA systems greatly facilitate the information access of users, which are popularly used in both Web and mobile Internet. In these systems, users input a question either by text or voice and expect to get the right answer quickly. However, due to factors such as thoughtless questions from users, misoperations of keyboard input and ASR (automatic speech recognizer) error, the ill-formed questions asked by users are usually expressed with vagueness, ambiguity, noises and errors.

By manual analysis on the WikiAnswer dataset<sup>1</sup>, we find that about 68% questions are ill-formed. As shown in Table 1, there are three typical ill-formed question types, specifically, wrong words, ill words order and noisy background, and they include 79%, i.e., ((21% + 23% + 12%)/68%) ill-formed questions. Generally, a question is a short sentence with a few words in QA systems. Directly using ill-formed questions to search for answers in a retrieval based QA systems [9] will hurt the downstream steps, e.g., answer selection [33] and hence compromise QA systems' effectiveness.

Inspired by the task of query refinement in web search [22], we study the task of question refinement in QA system, which aims to improve the quality of users' questions, in the meanwhile, boost the accuracy of the downstream answer retrieval. We can see that the task is complex since it contains the following subtasks: 1) word correction, e.g., correct "defenition" to "definition"; 2) word reorder, e.g., ill words order example in Table 1; 3) sentence simplification, e.g., remove the redundant expression like "based on timesteps prediction" in noisy background example in Table 1.

An intuitive way is to tackle these problems one by one alone. For instance, Xie et.al [37] proposed a character-level text correction

<sup>1</sup><http://knowitall.cs.washington.edu/oqa/data/WikiAnswers>

Type of Ill-formulation	Ratio in Data	Ill-formed Question	Well-formed Questions
Wrong words	21%	what is the <u>defenition</u> of the word infer	what is the <u>definition</u> for inference
Ill words order	23%	limestone is what kind of <u>rocke</u>	what is limestone <u>rock</u>
Noisy background	12%	based on tiresias prediction which heroic qualities will odysseus need to rely upon as he continues his journey	what heroic qualities does odysseus rely on

**Table 1: Examples of the three common types of ill-formed and generated well-formed questions on WikiAnswers dataset. The ratio in data is counted within 1000 random sampled triples. The other ill-formed questions belong to several other types which have a minority percentage.**

to deal with the orthographic errors and rare words. Yuan et.al [38] focus on grammar error correction to correct the erroneous word phrases. Besides, Zhang et.al. [39] utilized deep reinforcement learning to simplify questions, like splitting complex questions and substitutes difficult words with common paraphrases. However, it's laboursome to combine these methods together in practice, which might require no domain knowledge and a few human intervention.

Is it possible to tackle these problems with a unified model? Inspired by the successful usage of sequence-to-sequence (Seq2Seq) model [31] on related tasks such as machine translation [5], text summarization [21], and sentence simplification [39], it is promising to use it in the question refinement task. Seq2Seq model is flexible enough to encode patterns for sequence transformation such as word correction, word recorder, and sentence simplification, if there are appropriate training datasets. Unfortunately, we find that the vanilla Seq2Seq model does not perform well on this task. The reasons may be twofold: 1) it fails to learn a good representation of ill-formed questions, which might contain many wrong or noisy words. 2) The maximize likelihood objective is not consistent with our target, i.e., generated better quality questions and thus improve the accuracy of answer retrieval.

To overcome these problems, we develop a Seq2Seq model for question refinement called QREFINE. For the question representation, since a well-formed question might sensitive to the word order, we make use of the recent proposed contextual word embeddings such as BERT [7] to capture the contextual word information. As BERT is trained over a large scale unlabeled corpus, it also can alleviate the data sparsity problem where there is not enough training data. Moreover, considering the ill-formed questions might contain typos, we also incorporating the fine-grained character embedding [24] as a part of question representation. Our experimental results show that the two types of representations substantially improve the effectiveness of the Seq2Seq model.

To make the Seq2Seq model generate desired questions, we develop a training algorithm based on reinforcement learning. we assign not only word-level rewards to each word for its wording from a pertained language model and Bert language model as immediate rewards but also question-level rewards such as the correlation of the refined question to its answer. In order to solve the low data efficiency and unrobust policy problems on the traditional policy gradient method, we use advanced policy gradient method proximal policy optimization (PPO) [34] for well-formed question generation [29, 34]. We compared our model with the state-of-the-art baselines in two QA datasets. The result shows our model outperforms

baselines on question refinement. Besides, the case studies show the improved readability of the questions after refinement using QREFINE, and its effectiveness in improving the utility of an existing QA system. Moreover, it's worth to notice that our model is fully data-driven and might not require domain knowledge and human intervention.

## 2 PRELIMINARY

We formally define the question refinement task studied in this paper. After that, we introduce some terminologies that we will use throughout the paper.

### 2.1 Problem Description

Given an ill-formed question consists of  $\mathbf{x} = [x_1, x_2, \dots, x_N]$  of an arbitrary-length  $N$ , the well-formed question  $\mathbf{y} = [y_1, y_2, \dots, y_M]$  of a variable-length  $M$ . The aim of question refinement is to refine  $\mathbf{x}$  to  $\mathbf{y}$  which has better readability. It is expected that the generated well-formed question  $\mathbf{y}$  can be better able to retrieve the best answer candidate  $\mathbf{a}_k = [a_1, a_2, \dots, a_L]$ , where  $1 \leq k \leq s$  from an answer candidate pool  $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_s\}$ .

### 2.2 Seq2Seq Framework on Question Refinement

The Seq2Seq model adopts an encoder-decoder framework that learns to encode an ill-formed question  $\mathbf{x}$  into a fixed-length vector representation and to decode the fixed-length vector back into a variable-length well-formed question  $\mathbf{y}$ . From a probabilistic perspective, Seq2Seq model is a general method that learns the conditional distribution over a variable-length sequence conditioned on another variable-length sequence, namely,  $p_{lm}(y_1, \dots, y_M | x_1, \dots, x_N)$ .

The encoder can be a convolution neural network or a recurrent neural network that summarizes the ill-formed question into a vector representation. Since LSTM [13] is good at learning long-term dependencies in the data [11], we adopt LSTM to sequentially encode each word of ill-formed question  $\mathbf{x}$ . As the LSTM reads each word, the hidden state of the LSTM is updated  $h_n = \text{LSTM}_{\text{encoder}}(h_{n-1}, x_n)$ . Therefore, the encoder transforms the ill-formed question  $\mathbf{x}$  into a sequence of hidden states  $(h_1, h_2, \dots, h_N)$ .

The decoder can be another LSTM which is trained to generate the current hidden state  $k_m$  based on the current word  $y_m$  and the previous hidden state  $k_{m-1}$ :

$$k_m = \text{LSTM}_{\text{decoder}}(k_{m-1}, y_m). \quad (1)$$

Moreover, as introduced in [18], a context vector  $c_m$  can be obtained for each decoder step  $m$  by being attentive to the encoding of the source question dynamically:

$$c_m = \sum_{n=1}^N \alpha_{nm} h_n, \quad (2)$$

where  $c_m$  is a weighted sum of the hidden states of the ill-formed question  $\mathbf{x}$ : The attention score  $\alpha_{nm}$  between the  $n$ -th ill-formed question hidden unit  $h_n$  and  $m$ -th well-formed question hidden unit  $k_m$  is calculated as follows:

$$\alpha_{nm} = \frac{\exp(h_n^T \cdot k_m)}{\sum_{l=1}^N \exp(h_l^T \cdot k_m)}. \quad (3)$$

Formally, the Seq2Seq model is formed as below:

$$p_{lm}(y_m | y_{1:m-1}, \mathbf{x}) = \text{softmax}(g(k_m^T, c_m)), \quad (4)$$

where  $g(\cdot)$  is an activation function  $g(k_m^T, c_m) = \mathbf{W}_o \tanh(\mathbf{U}_h k_m^T + \mathbf{W}_h c_m)$ , where  $\mathbf{W}_o \in \mathbb{R}^{|V| \times d}$ ,  $\mathbf{U}_h \in \mathbb{R}^{d \times d}$  and  $\mathbf{W}_h \in \mathbb{R}^{d \times d}$ ;  $|V|$  is the output vocabulary size and  $d$  is the hidden unit size.

### 3 REINFORCED GENERATIVE QUESTION REFINEMENT

#### 3.1 Model Description

Despite the successful application in numerous sequence transduction tasks [2], a vanilla Seq2Seq model is not ideal for question refinement since it only makes a few trivial changes of ill-formed question [39]. To encourage a wider variety of rewrite operations while keeping the refined question fluent and coherent to the answer, we employ a reinforcement learning framework (see Figure 1).

The refinement agent first reads the ill-formed question  $\mathbf{x}$  from the encoder; and then at each step of the decoder, it takes an action  $y_m \in V$ , where  $V$  is the output vocabulary, according to a policy  $\pi_\theta(y_m | y_{1:m-1}, \mathbf{x})$ . The agent continues to take actions until it produces  $\langle \text{EOS} \rangle$  (denoting end of sentence) token yielding the generated well-formed question of our model  $\mathbf{y} = [y_1, y_2, \dots, y_M]$ . Two types of rewards, wording reward, and answer correlation reward, are received and the advanced policy gradient method PPO is used to update the agent. In the following, we introduce our question representation and reward function. After that, we present the details of the processes for generating accurate and consistent questions by leveraging the REINFORCE and PPO method.

#### 3.2 Question Representation

The ill-formed question has a wrong semantic order and contains some unrelated background information. To solve the problem, the model need to learn the institutional utterance of words, which needs to consider the correlation between words and words. The widely-used context-free models such as Skip-gram [19] or GloVe [26] cannot consider the correlation between words. Because they generate a single word embedding representation for each word in the vocabulary, so “apple” would have the same representation in “red apple” and “apple store”. However contextual models, like BERT can generate a representation of each word based on the

other words in the sentence. Therefore, we concatenate the context-free and contextual embedding together as the word embedding to capture such coarse-grained correlation patterns of words.

Since the ill-formed question always contains the misspelled words, which is usually set as  $\langle \text{UNK} \rangle$ , which is hard to capture the meaning of the original word. To capture the meaning of the misspelled words, we extend the word expression by incorporating fine-grained character expression. By using the character-level embedding, we can get the high-dimensional vector representation. As character-level input, the original sentence is decomposed into a sequence of characters, including special characters, such as quotation mark. Characters are embedded into vectors, which can be considered as 1D inputs to the Bidirectional Long Short-term Memory Network (BI-LSTM) [13], and the hidden layer of the last LSTM unit is the fixed-size vector of each word. Overall, we combine the fine-grained character-level embedding and coarse-grained contextual and context-free word embedding to represent the question.

#### 3.3 Reward

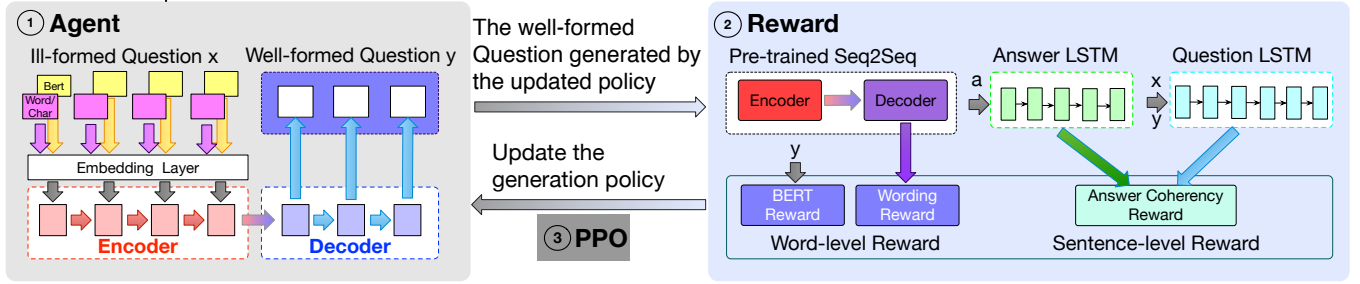
The reward for system output  $\mathbf{y}$  is the weighted sum of two types of rewards aimed at achieving well-formed readable and answer correlation question: wording rewards on the word-level from the Reward RNN and BERT, which aims to measure how well each generated word is in line with the language model (LM) rule, and the question-level answer correlation reward that has the ability to infer the correlation of the refined question to its answer, even if it is not generating until the end of the well-formed question.

**Wording Reward** The wording reward  $r_w$  aims to give an immediate reward to each of words when it is being generated in the well-formed question. BERT pre-trained on the large dataset like Wikipedia could give the contextual wording reward  $r_B(y_t) = p_B(y_t | y_1, \dots, y_M)$ , which is the probability of word  $y_t$  given by BERT model. Moreover, for the domain-specific, we also use the decoder of the pre-trained Seq2Seq module as a trained LM Reward RNN which is able to score the probability of the next word given the words generated so far. Thus, the wording reward of the  $t$ -th word in the well-formed question is:

$$r_w(y_t) = r_B(y_t) + p_{lm}(y_{t+1} | k_t), \quad (5)$$

where  $k_t$  is the current state which is the hidden representation of the generated well-formed question with  $t$  words so far:  $[y_1, y_2, \dots, y_t]$  and  $y_{t+1}$  is the generated word in the  $(t+1)$ -step.

**Answer Correlation Reward** The refinement result should not only improve the readability of the question, more importantly, have better ability to address its correlation to the answer once refined. With this motivation, we design an answer correlation reward  $r_{ac}$  to further measure the correlation of the refined question to its answer on the question-level as a whole. As answers themselves are sometimes ill-formed and contain a large amount of unrelated information, they may not share lexical units with the well-formed question directly. But the well-formed question is semantically easier to answer than the ill-formed question. Following the similar ranking loss in [10, 33], the answer correlation module defines the training objective as a hinge loss:



**Figure 1: The architecture of the proposed model QREFINE.** ① The encoder of the agent module reads the ill-formed question and the decoder generates a well-formed question, one word/phrase at a time. ② The well-formed question being generated so far is sent to a pre-trained reward module, which calculates a word-level wording reward from word-level LM and BERT Reward and a question-level answer correlation reward from QA similarity. ③ The PPO module updates agent’s generation policy, aiming to maximize the rewards.

$$r_{ac}(y) = \max\{0, \epsilon - \text{sim}(\text{LSTM}_q(\mathbf{x}), \text{LSTM}_a(\mathbf{a})) + \text{sim}(\text{LSTM}_q(\mathbf{y}), \text{LSTM}_a(\mathbf{a}))\}, \quad (6)$$

where we use two separate LSTMs,  $\text{LSTM}_q$  and  $\text{LSTM}_a$ , to encode the question and answer to the vector representation. The well-formed question and ill-formed question share the same  $\text{LSTM}_q$ , and  $\epsilon$  is the constant margin. Furthermore,  $\text{sim}(\text{LSTM}_q(\mathbf{x}), \text{LSTM}_a(\mathbf{a})) = \text{LSTM}_q(\mathbf{x})\mathbf{W}_{\text{sim}}\text{LSTM}_a(\mathbf{a})^T$  computes a bi-linear term between the question and its correct answer. We train the model by maximizing answer correlation reward  $r_{ac}$  using ground-truth well-formed and ill-formed questions to learn the weight of  $\text{LSTM}_q$ ,  $\text{LSTM}_a$  network and  $\mathbf{W}_{\text{sim}}$ . After that, we hold a fixed copy of networks to give rewards to the generated well-formed question.

**Accumulated Reward** We add the answer correlation reward to the end of the wording reward, as the overall evaluation of the generated question. The QREFINE reward  $r$  of each word is the combination of the wording reward and the answer correlation reward,

$$r(y_i) = \begin{cases} r_w(y_i), & i \neq M \\ r_w(y_i) + c_1 r_{ac}(\mathbf{y}), & i = M \end{cases} \quad (7)$$

where  $c_1$  is the parameter to tune the weight between wording reward  $r_{lm}$  and answer correlation reward  $r_{ac}$ ; Since we want QREFINE module to have the ability to infer reward even if not reaching the end of the generation process and the future reward will influence the current reward, we adopt the accumulated QREFINE reward  $R$  with the discounted factor  $\gamma$  and the accumulated QREFINE discounted reward of  $t$ -th word is represented as,

$$R(y_t) = \gamma^0 r(y_t) + \gamma^1 r(y_{t+1}) + \dots + \gamma^{M-t} r(y_M). \quad (8)$$

By using the accumulated reward, we are able to infer the answer correlation reward even if we are not reaching the end of the generation process.

### 3.4 Question Generation

A popular choice of loss in traditional models is the cross-entropy used to maximize the probability of the next correct word. However, this loss is at the word-level and the performance of these models

is typically evaluated using discrete metrics. To overcome, we draw on the insights of deep reinforcement learning, which integrates exploration and exploitation into a whole framework. Instead of learning a sequential recurrent model to greedily look for the next correct word, we utilize a policy network and a reward function to jointly determine the next best word at each time step, which aims to maximize the reward of the whole sentence.

Question refinement can be formulated as a Markov decision process (MDP)  $(S, A, P, R)$ , where  $S$  is a set of states  $s_t = \{y_{1:t-1}, \mathbf{x}\}$ ,  $A$  is a set of actions  $a_t = y_t$ ,  $P$  is the transition probability of the next state given the current state and action,  $R$  is a reward function  $r(s_t, a_t)$  for every intermediate time step  $t$ , and  $\gamma$  is a discount factor that  $\gamma \in [0, 1]$ .

The actions are taken from a probability distribution called policy  $\pi$  given the current state (i.e.,  $a_t \sim \pi(s_t)$ ). In question refinement,  $\pi$  is a seq2seq model. Therefore, reinforcement learning methods are suitable to apply to question refinement model by learning the seq2seq model, or policy  $\pi$ , that can gain reward as much as possible.

**3.4.1 On-policy Optimization.** Due to the high dimensional action space for question refinement and high diversity of the required generation result, policy gradient method, like REINFORCE [4] are more appropriate in the question generation than value-based methods like Q-learning [20].

For a given ill-formed question  $\mathbf{x}$ , we want to return a formulated question  $\mathbf{y}$ , maximizing an accumulated reward  $R$ . The answer  $\mathbf{a}$  is the known given by the database. The question  $\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})$  is generated according to  $\pi_\theta$  where  $\theta$  is the policy’s parameter and the goal is to maximize the expected reward of the reformulated question under the policy,  $E_{\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})}[R(\mathbf{y})]$ .

Given reward  $r_t$  at each time step  $t$ , the parameter  $\pi$  of policy  $\pi$  (a seq2seq model) is updated by policy gradient as follows:

$$E_{\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})}[R(\mathbf{y})] \approx \frac{1}{N} \sum_{i=1}^N R(y_i), y_i \sim \pi_\theta(\cdot|\mathbf{x}). \quad (9)$$

To compute gradients for training we use REINFORCE[36],

$$\begin{aligned} & \nabla_{E_{y \sim \pi_\theta(\cdot|\mathbf{x})}}[R(y)] \\ &= E_{y \sim \pi_\theta(\cdot|\mathbf{x})} \nabla_{\log(\pi_\theta(y|\mathbf{x}))} R(y) \\ &\approx \frac{1}{N} \sum_{i=1}^N \nabla_{\log(\pi_\theta(y_i|y_{1:i-1}, X))} R(y_i), y_i \sim \pi_\theta(\cdot|\mathbf{x}). \end{aligned} \quad (10)$$

This estimator is often found to have high variance, leading to unstable training [12]. We reduce the variance by using a baseline:  $B(\mathbf{x}) = E_{y \sim \pi_\theta(\cdot|\mathbf{x})}[R(y)]$  [32]. This expectation is also computed by sampling from the policy given  $\mathbf{x}$ .

To avoid the model not being able to explore new words that could lead to a better answer correlation question, we use entropy regularization:

$$\mathbf{H}[\pi_\theta(y|\mathbf{x})] = \sum_{i=1}^N \sum_{y_i \in V} \log(\pi_\theta(y_i|y_{1:i-1}, \mathbf{x})) \pi_\theta(y_i|y_{1:i-1}, \mathbf{x}) \quad (11)$$

The final objective is:

$$\mathbb{E}_{y \sim \pi_\theta(\cdot|\mathbf{x})} [R(y) - B(\mathbf{x})] + \lambda \mathbf{H}[\pi_\theta(y|\mathbf{x})], \quad (12)$$

where  $\lambda$  is the regularization weight.  $R(y) - B(\mathbf{x})$  can be interpreted as the goodness of adopted action at over all the possible actions at state  $\mathbf{x}$ . Policy gradient directly updates  $\pi$  to increase the probability of at given  $s_t$  when advantage function is positive, and vice versa.

**3.4.2 Off-policy Optimization.** The vanilla policy gradient method is on-policy method and have convergence problem [28]. Empirically they often lead to problems like low data efficiency and unreliable performance, as shown in subsection 4.6. We use the advanced deep reinforce method proximal policy optimization (PPO) [29] to learn a more stable policy.

Proximal policy optimization (PPO) [34] is an approximation method of trust region policy optimization (TRPO) [28]. Different from TRPO which uses a second-order Taylor expansion, PPO uses only a first-order approximation, which makes PPO very effective in RNN networks and in a wide distribution space:

$$\max_{\theta} L^{TRPO}(\theta) = \mathbb{E} \left[ \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_t \right], \quad (13)$$

$$\text{subject to } \mathbb{E}[\mathbf{KL}[\pi_{\theta_{old}}(a_t|s_t), \pi_\theta(a_t|s_t)]] \leq \delta$$

where  $\pi_{old}$  is the old parameters before update. The updated policy  $\pi$  cannot be too far away from the old policy  $\pi_{old}$ , because the KL-divergence between  $\pi$  and  $\pi_{old}$  is bounded by the small number  $\delta$ .

To optimize policy, PPO alternates between sampling sentence generated from the current policy and performing several epochs of optimization on the sampled sentences. According to the paper [29], the clipped objective to heuristically constrain the KL-divergence setting achieves the best performance:

$$L_t^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta) \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)) \hat{A}_t], \quad (14)$$

where  $\beta_t$  denotes the probability ratio  $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$  and  $\epsilon$  is a hyperparameter (e.g.,  $\epsilon = 0.1$ ). When  $\hat{A}_t$  is positive, the objective is clipped by  $(1 + \epsilon)$ ; otherwise the objective is clipped by  $(1 - \epsilon)$ .

---

#### Algorithm 1 QREFINE-PPO

---

**Input:** Ill-formed question  $X$ , Well-formed question  $Y$ , rating data  $R$ , the number of episodes  $K$ ,  $\epsilon$ -greedy parameter  $\epsilon$ , ratio of language model reward and answer correlation reward  $c_1$ , the discounted factor of RL  $\lambda$ , the threshold  $\epsilon$  and the entropy regularization  $c_2$

**Output:** the learned policy  $p_\theta$

- 1: Initialize policy  $p_\theta$  and old policy  $p_{\theta_{old}}$  with supervised pre-trained policy  $p_{\theta'}$
  - 2: **for** episode = 1, ...,  $K$  **do**
  - 3:   Uniformly pick a batch of ill-formed question  $u \in \mathcal{U}_{train}$  as the environment
  - 4:   Start to generate the word according to  $p_{\theta_{old}}(y_i|X)$  until the <EOS> token is generated, the generated sentence as  $Y'$
  - 5:   Send  $X$  and  $Y'$  to the BERT mechine and pretrained word embedding model to calculate the word-level reward
  - 6:   Send  $X$ ,  $Y$  and  $Y'$  to the qa-lstm model to calculate the sentence-level reward, based on Eq. 7
  - 7:   Calculate the advantage function of each time step according to Eq. 15
  - 8:   **repeat**
  - 9:     Update the policy  $p_\theta$  using Eq. 16
  - 10:   **until** convergence
  - 11:   Set old policy  $p_{\theta_{old}}$  to policy  $p_\theta$
  - 12: **end for**
- 

$\hat{A}_t$  is the expected advantage function (the expected rewards minus a baseline like value function  $V(k_t)$  of time  $k_t$ ) which can be calculated as:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \dots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \quad (15)$$

$$\delta_t = r_t + \gamma V(k_{t+1}) - V(k_t).$$

To improve the exploration of our model for generating diverse yet coherent words that could constitute a better well-formed question, we use entropy regularization. The integrated PPO method is shown as below:

$$L_t^{PPO}(\theta) = \mathbb{E}_t[L_t^{CLIP}(\theta) + c_2\pi_\theta(y_{t+1}|k_t)\log(\pi_\theta(y_{t+1}|k_t))]. \quad (16)$$

The algorithm of QREFINE with PPO optimization shows in Alg. 1

**Learning** The training stage of traditional models suffer from the exposure bias [27] since in testing time the ground-truth is missing and previously generated words from the trained model distribution are used to predict the next word. In the testing phase, this exposure bias makes error accumulated and makes these models suboptimal, not able to generate those words which are appropriate but with low probability to be drawn in the training phase.

In order to solve the exposure bias problem, we train the model by using MIXER algorithm described in [27] to expose both training data and its predictions. In the inference stage, we greedily selected the word that has the highest probability to generate the question stopping until the <EOS> token is generated.

## 4 EXPERIMENTS

In this section, we evaluate the proposed methods on two real-world datasets, by examining the readability of the refined questions

both quantitatively and qualitatively. A question-answer retrieval experiment is used to evaluate the performance of the generated question. In the end, we further conduct the ablation study and learning curve of the method.

#### 4.1 Dataset

The datasets are formatted as triples, where each instance consists of a well-formed question, an ill-formed question and an answer in each triple.

**Yahoo:** The Yahoo non-factoid question dataset<sup>2</sup> is collected from Yahoo Webscope that the questions would contain non-factoid answers in English. After limiting the length of the answer, we have 85k questions and their corresponding answers. Each question contains its best answer along with additional other answers submitted by users. The average length of ill-formed questions is around 12 tokens and the average length of well-formed questions is around 10 tokens, with 73k vocabulary size and 44 different characters. The average length of answers is around 39 tokens.

To testify the refinement performance from ill-formed to well-formed question, we generate the ill-formed question on three types of the ill-formed question, wrong words, wrong order and Noisy background. We randomly change the character of the words or change the order of the character of words to generate the **Wrong Word** dataset. For the **Wrong Order** dataset, we randomly change the order of some fragments in the well-formed question. For the **Noisy Background** dataset, we randomly sampled an arbitrary length phrase from the other answer and add to the original clean question. For the **Yahoo** dataset, we randomly execute those three operations to generate threefold noisy questions, which contains 254k triples. The example of ill-formed question in those datasets are shown in Table 2.

**Customer Service Userlog (CSU):** This anonymized dataset contains online question-answering Userlog from a commercial customer service platform containing 1 million instances in Chinese language. The ill-formed question is the question asked by users and the well-formed question is selected from a pool of FAQs collected by editors. After we delete the duplicated triples, the left triple size is 111k. The average length of ill-formed questions is around 6 tokens, and the average length of well-formed questions is also around 6 tokens while the average length of answers is around 54 tokens with 14k vocabulary size and 2041 characters.

#### 4.2 Baselines and Benchmarks

The compared methods are summarized as follows:

**Seq2Seq** is a basic encoder-decoder sequence learning system with Luong attention [18] and Bi-direction LSTM on encoder model. **PARA-NMT** [8] is a NMT-based question paraphrasing method which assigns higher weights to those linguistic expressions likely to yield correct answers. **AQA** [6] is the reinforce method seeking to reformulate questions such that the QA system has the best chance of returning the correct answer in the reading comprehension task. Since our datasets do not contain the context information, we use the QA-lstm to measure the similarity between the generated question and the answer as the reward. Following [5], we use a

<sup>2</sup><https://ciir.cs.umass.edu/downloads/nfL6/>

bidirectional LSTM as the encoder and a 4-layer stacked LSTM with attention as the decoder. **TOQR** [23] is the query reformulation method with reinforcement learning to maximize the number of relevant documents returned. TOQR use reinforcement method to select terms from the original query and candidate retrieved document to reformulate the query, and the reward is the document recall.

Since the proposed QREFINE consists of several components, we consider several variations of QREFINE as follows:

**QR-word** is the reinforce model only using wording reward, which is viewed as word-level reward. **QR-ans** is the reinforce model using answer correlation as the reward, which is views as question-level reward. **QREFINE-RF** combines both word-level wording reward and question-level answer correlation reward and uses REINFORCE policy gradient based to optimize.

**QREFINE-PPO** is the proposed model using PPO and combining both word-level wording reward and question-level reward.

**Experimental Setting** We randomly divide the dataset into a training | development | test set (80% | 10% | 10%). We tune the hyperparameters on development set and report results on test set. We implement all model by Tensorflow using Python on a Linux server with Intel Xeon E5-2620 v4 CPU and an NVIDIA 1080Ti GPU. The code is available on Github.<sup>3</sup>

On Yahoo dataset, we use the released skip-gram model word embedding [19] with 300 dimensions<sup>4</sup>. We fix the word representations during training. The character embedding of is 50 dimensions for each character. The number of hidden unit in character Bi-LSTM is 100. We add <EOS> at the end of sentences. And we set the word out of the vocabulary as <UNK>. We choose word embedding of 200 dimensions for CSU dataset and use the Glove model [26] to get the pre-trained word embedding. The character embedding of CSU dataset is 50 dimensions for each character. The number of hidden unit in character Bi-LSTM is 50. For BERT word embedding<sup>5</sup>, it gives us 768 dimensions of the word embedding on both datasets. We combine contextual-free word embedding, BERT embedding and character embedding for each of word on both dataset.

We set the LSTM hidden state to 300 on CSU dataset and 500 on WikiAnswer dataset. Optimization is performed using Adam [14], with an initial learning rate of 0.001. The mini-batch size for the update is set at 64 on both datasets. During reinforcement training, we set the ratio of language model reward and answer correlation reward  $c_1$  as {0.1, 1, 10}, the discounted factor of RL  $\lambda$  is [0, 1), the threshold  $\epsilon$  is {0.1, 0.2, 0.3} and the entropy regularization  $c_2$  is {0.1, 1}.

#### 4.3 Question Generation

In this section, we give the experimental analysis to quantitatively and qualitatively evaluate the quality of generated questions.

**4.3.1 Quantitative Evaluation of Question Generation.** To evaluate the quality of the generated well-fined question, we first use automatic metrics to quantitatively show the performance. We

<sup>3</sup><https://github.com/yeliu918/QREFINE-PPO>

<sup>4</sup><https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md>

<sup>5</sup><https://github.com/hanxiao/bert-as-service>

Original well-formed question : why don't european restaurants serve water?	
Type	Ill-formed question
Wrong word/typo	why don't <i>oeurpan rantaurest</i> serve <i>wataar</i> ?
Wrong order	european restaurants why serve don't water?
Noisy background	concerned with the digestive process why don't european restaurants serve water?
Three operations	why <i>digistive</i> process with the <i>restarunts</i> european don't serve water?

Table 2: Example of the three operations to generate ill-formed question on Yahoo Dataset

Method	Wrong Word			Wrong Order			Noisy Background		
	BLEU-1	Rouge	Meteor	BLEU-1	Rouge	Meteor	BLEU-1	Meteor	Rouge
Seq2Seq	47.10	60.41	30.95	53.11	67.67	36.37	50.17	62.19	31.75
PARA-NMT	53.10	64.91	35.59	59.13	73.75	41.27	55.37	68.45	37.56
TOQR	43.15	56.04	28.47	49.60	45.49	56.76	32.75	57.85	45.39
AQA	61.93	77.36	45.91	63.34	80.83	50.15	61.08	74.14	42.10
QREFINE-RF	67.82	83.16	51.07	70.74	87.21	55.24	69.12	85.21	53.17
QREFINE-PPO	<b>68.83</b>	<b>84.76</b>	<b>52.72</b>	<b>72.22</b>	<b>88.94</b>	<b>56.22</b>	<b>71.57</b>	<b>86.12</b>	<b>53.22</b>

Table 3: Question Generation Evaluation on Yahoo dataset to test models ability to correct wrong words, order and remove background.

Method	Yahoo						CSU					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge	Meteor	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge	Meteor
Seq2Seq	39.50	31.53	23.81	16.78	53.07	22.76	72.76	53.17	37.49	26.17	68.39	35.57
PARA-NMT	41.08	33.74	26.50	19.73	55.10	23.80	73.18	59.41	56.53	36.44	69.05	57.18
AQA	43.40	37.14	30.80	24.58	58.37	27.17	74.13	65.53	58.00	31.63	74.40	62.67
TOQR	31.23	20.69	12.45	5.89	41.92	15.19	48.92	44.69	40.34	35.73	65.10	33.00
QR-word	47.73	42.46	37.09	31.46	63.03	31.00	77.49	69.61	62.00	34.91	77.71	66.17
QR-ans	47.17	41.80	36.36	30.66	62.48	30.55	78.50	70.95	63.08	36.99	78.72	67.07
QREFINE-RF	48.72	44.20	39.58	34.74	64.60	32.59	79.71	72.43	64.59	37.54	79.96	68.40
QREFINE-PPO	<b>50.90</b>	<b>47.47</b>	<b>43.91</b>	<b>40.19</b>	<b>67.41</b>	<b>35.37</b>	<b>82.55</b>	<b>75.54</b>	<b>67.63</b>	<b>40.33</b>	<b>82.74</b>	<b>71.54</b>

Table 4: Question Generation Evaluation on Yahoo and CSU dataset.

use the precision-based automatic metrics BLEU-1, BLEU-2, BLEU-3, BLEU-4 [25] which measures the average n-gram precision on a set of reference sentences, with a penalty for overly short sentences, and ROUGE [17] based on recall and METEOR [3] that is based on both precision and recall to measure the generation results.

In Table 3, our model performs best on each single task, wrong word, wrong order or removing noisy background. Since QREFINE uses character-level embedding, contextual-free and contextual word embedding BERT, it can better deal with the misspellings and understand the ill-formed question than all baselines. By using the word-level reward, our model can learn a better language policy, hence it can perform well on correcting the word order and wrong word. Besides, since our model considers the answer correlation as the reward, so it can capture the useful information in the question and achieves superior results on the noisy background.

For the three operations composite task, as shown in Table 4, our model performs the best on both datasets. In PARA-NMT, the paraphrased questions are very similar, and therefore there is a great chance that paraphrased questions always get the same answer. The reinforced query reformulation method TOQR reformulates the query by selecting terms from the retrieved documents by using the document recall as reward. Since the query is a list of terms, therefore, the generated sequence of terms have poor readability, resulting in lowest performance among all baselines. The original

AQA refines the question for the reading comprehension. We use the QA-lstm as reward, which has similar framework as PARA-NMT but using REINFORCE optimization. Therefore AQA has the same problem as PARA-NMT. Compared to QR-ans, QR-word has a better performance, which shows that QR-word contributes more to the readability. QREFINE-PPO is higher than QREFINE-RF, which may because PPO method can improve the high variance problem using REINFORCE method on question refinement task.

**4.3.2 Case Study.** To demonstrate the readability and effectiveness of the proposed method, Table 5 shows examples of generated outputs on Yahoo and CSU dataset. As we can see, QREFINE has the ability to correct the wrong words (shown in Case 1), change the question word order for better readability (shown in Case 2) and remove the unnecessary background in the original question (shown in Case 3). For example, “differncee” can be correct as “difference”. And the refined question by our model is readable. But the question generated by other baselines cannot well express the original meaning in the ill-formed question or be misleading and also have problems like repeatedly the useless words (e.g., Seq2Seq in Yahoo/Case 1), no natural language sentence (e.g., TOQR in Yahoo/Case 1) and express the different meaning with the ill-formed question (e.g., AQA in Yahoo/Case 1). Therefore, the question generated by QREFINE is more readable than other alternatives and is able to keep the original user intention of asking the question.

Yahoo		
Dataset/Case	Ill-formed	Well-formed
Yahoo/Case 1 Wrong word	what's the <u>differncee</u> between climate <u>chnage</u> and global <u>wraming</u> ?	Seq2Seq: what's the <u>difference</u> between human beings and cancer cancer?
		TOQR: <u>differncee</u> between climate <u>chnage</u> and global <u>wraming</u> oxygen weather <u>warming</u>
		AQA: what's the <u>difference</u> between global and global <u>warming</u> war?
		QREFINE: what's the <u>difference</u> between climate <u>change</u> and global <u>warming</u> ?
Yahoo/Case 2 Wrong order	how safe to meet is it an american?	Seq2Seq: how to make it is an adult professional?
		TOQR: how safe meet an american again powerful secretary
		AQA: an animal how safe is it to meet an american?
		QREFINE: how safe is it to meet an american?
Yahoo/Case 3 Noisy Background	who would able help what is string theory and how is it used?	Seq2Seq: what is empirical and how is it used mechanics?
		TOQR: able help string theory and how is it used theory physics used model
		AQA: what is turbo motion and how it is used?
		QREFINE: what is string theory and how is it used?
CSU		
Dataset/Case	Ill-formed	Well-formed
CSU/Case 1 Wrong word	??????? ????? I want to ask why <u>coubon</u> cannot be used	Seq2Seq: ?????? How to use the <u>coupon</u>
		TOQR: ??? ??? ??? ??? <u>coubon</u> <u>coupon</u> service regulations validity bargain goods
		AQA: ?????? how can I get coupon
		QREFINE: ?????????? What should I do if the <u>coupon</u> cannot be used
CSU/Case 2 Wrong order	???????? I do what phone has water damage	Seq2Seq: ?????? how do I find the lost phone
		TOQR: ??? ???? ?? water damage avoid water absorb water
		AQA: ????? my phone has water damage
		QREFINE: ?????? What should I do if the phone has water damage
CSU/Case 3 Noisy Background	????? ??????? forgot the password and repeatedly log in right now cannot log in	Seq2Seq: ?????? What should I do when cannot log in
		TOQR: ??? ???? ???? ???? forget password cannot login block account number wrong password
		AQA: ?????? What should I do when I forget password
		QREFINE: ?????????? What should I do when login password forgot

Table 5: Cases study of Generated Results on Yahoo dataset and Commercial Customer Service Userlog dataset. Typos and substitutions are shown in underscore.

#### 4.4 Answer Retrieval

To validate the effectiveness of question refinement in helping retrieve answers in existing QA systems, we use PyLucene<sup>6</sup> for retrieving the answer to the search question.

**Hits@K:** The top K relevant answers retrieved by the search question using PyLucene. If the gold answer is inside of the top K retrieved answers, then the Hits@K of this search question equals 1, otherwise, it equals 0. The whole Hits@K of questions is the average development set question's Hits@K.

**Results** In Table 6, we can see the all refined question generated can be better than the Ill-formed question, which shows that the refinement process is needed. TOQR, which aims to maximize the retrieval results achieves the good performance compared with other methods. However, our model QREFINE achieves very better performance comparing with TOQR in most case and the question refined by our model has better readability. The Hits@K score retrieved by QREFINE-word is higher than Seq2Seq, which indicates that by improving the readability of question, the retrieval ability also be improved. As QREFINE-sen performs better than Seq2Seq, it shows that the reward considering over-all question structure

for a better correlation with refined question to its answer is important. This result shows the superiority of QREFINE in greatly improving QA utility via explicitly refining the question for enhancing its readability and retrieve ability for both computer and human.

#### 4.5 Ablation Study

In order to find out which part of the model improves the automatic evaluation performance, we do the ablation study. **S2S+W** is seq2seq model using word-level embedding. **S2S+W&C** is seq2seq model using word-level and char-level embedding. **S2S+W&C&B** is seq2seq model using word-level, char-level, and BERT embedding. **QR-word** is seq2seq model considering three embeddings and using word reward to train the RL model. **QR-ans** is seq2seq model considering three embeddings and using answer coherence reward to train the model. **QR-RF** considers multi-grain word embedding and both word reward and sentence reward but uses REINFORCE method. **QR-PPO** is our model.

From Fig. 2(a) and 2(b), we can see that using multi-grain word embedding can help the model better to correct the ill-formed question than just using single word embedding. And using PPO reinforcement learning with the word-level and sentence-level reward

<sup>6</sup><http://http://lucene.apache.org/pylucene/>



Yahoo	Hits@1	Hits@3	Hits@5	Hits@10
Ill-formed	3.35	4.89	8.68	14.98
Seq2Seq	3.60	4.50	8.99	16.97
PARA-NMT	6.47	7.95	12.01	17.21
AQA	6.48	8.12	13.06	18.78
TOQR	7.27	10.25	22.83	<b>32.43</b>
QR-word	7.11	9.34	20.79	29.88
QR-ans	7.16	9.67	22.45	30.89
QREFINE	<b>8.09</b>	<b>10.76</b>	<b>23.95</b>	32.23

CSU	Hits@1	Hits@3	Hits@5	Hits@10
Ill-formed	10.34	18.72	26.78	39.78
Seq2Seq	11.84	19.83	27.25	40.12
PARA-NMT	15.59	21.09	30.95	40.33
AQA	16.89	20.00	31.81	40.67
TOQR	20.41	27.98	34.59	48.78
QR-word	21.23	25.98	33.04	46.19
QR-ans	19.64	26.11	34.67	47.76
QREFINE	<b>22.10</b>	<b>28.69</b>	<b>35.98</b>	<b>49.16</b>

Table 6: Answering Retrieval Result on Yahoo and CSU dataset.

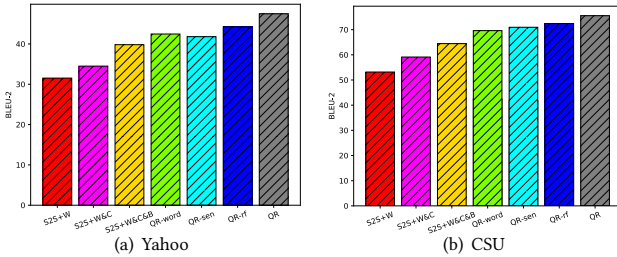


Figure 2: The ablation study on Yahoo and CSU dataset

can improve the model to learn a stable policy that can generate the appropriate well-formed question.

#### 4.6 Learning Curves Analysis

The BLEU-2 scores and learning curves of different optimization algorithms are presented in Figure 3(a) and 3(b). From the testing results in Table 4, we can see that the two optimization methods have comparable performance, but PPO achieves a slightly higher BLEU-2 score than REINFORCE. Moreover, we find out that the training progress of PPO is more stable and converge earlier than policy gradient. This shows that PPO methods can improve the high variance problem of using REINFORCE, and can help the learning converge quickly.

### 5 RELATED WORK

#### 5.1 Generative Text Refinement

The generative ability of deep neural networks leads to the prevalence of Seq2Seq models for reformulation task. These methods typically accomplish the reformulation by training a recurrent neural network such as an LSTM network to predict the next word in a sentence sequence. [23] uses the reinforcement learning to

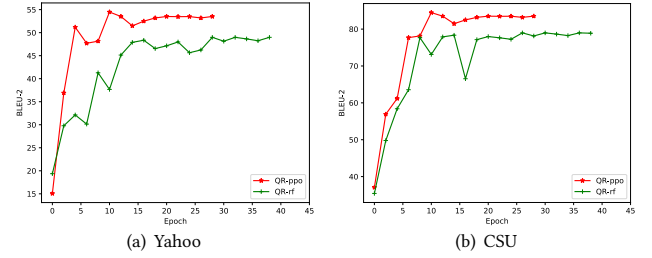


Figure 3: The learning curve analysis on Yahoo and CSU dataset

reformulate the query to maximize the number of relevant documents retrieved. Our work differs with their method in that we generate natural language sequences rather than terms; thus their reformulated query doesn't contain the propriety of readability and understandability. [39] proposed the sentence simplification task, which aims to make sentences easier to read and understand. Similarly, [37] operates on the character level to flexibly handle orthographic errors in spelling, capitalization, and punctuation. Although their frameworks can reform the sentences to be more readable, their objective does not include the capability of refining the question to get answers easier. Active QA (AQA) [6] uses an agent as a mediator between the user and a black box QA system, e.g. BiDAF, to generate questions that elicit the best possible answer. Since the pretrained fixed environment, BiDAF, is not updating with the model, feedback on the quality of the question reformulations could be quite noisy which presents a challenge for training. Moreover, BiDAF works on the reading comprehension, the answer is the paraphrase in the context, therefore there is a great change that this model always generates the same answer, which brings another challenge for training.

#### 5.2 Reinforcement Learning for QA

Due to the high dimensional action space for text generation and high diversity of the required generation result, policy gradient methods are more appropriate in the text generation than value-based methods like Q-learning [20]. By using policy gradient method, the limitation of cross-entropy loss that inherently comes with word-level optimization is alleviated and allowing sequence-level reward functions, like BLEU, to be used [27]. [1] extends this line of work using actor-critic training. Uses of policy gradient for QA include [16], who train a semantic parser to query a knowledge base, and [30] who propose query reduction networks that transform a query to answer questions that involve multi-hop common sense reasoning. Li et al. [15] use RL and SL to learn the paraphrase of the sentence. The on-policy method like REINFORCE suffers the high variance and slow to converge. The off-policy method like TRPO and PPO [28, 29] recently applied on the game like Atari. They can deal with the problems by regularizing the gradient of policy. Tuan et al [34] apply the off-policy gradient method to the sequence generation task and shows that PPO surpass policy gradient on stability and performance.

## 6 CONCLUSION AND FUTURE WORK

Question refinement aims to refine ill-formed questions, which typically includes various types of subtasks such as spelling error correction, background removal and word order refinement. Instead of tackle these subtasks separately, we develop a unified model, based on Seq2Seq, to handle this task in a data-driven way. We improve the question representation by incorporating character embedding and contextual word embedding such as BERT. To make the refinement process more controllable, we combine Seq2Seq model with deep reinforcement learning. We define a sequence generator by optimizing for a combination of imposed reward functions. The experimental results show that our method can not only produce more readable question but also significantly improves the retrieval ability of question for downstream QA system.

Question refinement is a challenging task and there are several directions to improve. One direction is to develop the advanced method, such as creating different awards that are more suitable to deal with the three subtasks. Besides, In our setting, the ill-formed and well-formed questions still need to be paired. In most of realistic cases, we only have a pool of well-formed. We seek to use inverse reinforcement learning [35] to learn the intrinsic representation of the well-formed question. Therefore, given an ill-formed question, the model can refine it to the well-formed. Finally, it is also interesting to make use of the result of question refinement to improve other related tasks such as question understanding and question answering recommendation.

## ACKNOWLEDGEMENTS

This work is supported in part by NSF under grants III-1526499, III-1763325, III-1909323, SaTC-1930941, and CNS-1626432.

## REFERENCES

- [1] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086* (2016).
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [3] Satandeep 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*. 65–72.
- [4] Andrew G Barto and Richard S Sutton. 1998. Reinforcement learning. *Neural systems for control* (1998), 7–29.
- [5] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906* (2017).
- [6] Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2018. Ask the Right Questions: Active Question Reformulation with Reinforcement Learning. *ICLR* (2018).
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [8] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022* (2017).
- [9] Manaal Faruqi and Dipanjan Das. 2018. Identifying Well-formed Natural Language Questions. *arXiv preprint arXiv:1808.09419* (2018).
- [10] Minwei Feng, Bing Xiang, Michael R Glass, Lidian Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 813–820.
- [11] Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).
- [12] Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research* 5, Nov (2004), 1471–1530.
- [13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [15] Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2017. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279* (2017).
- [16] Chen Liang, Jonathan Berant, Quoc Le, Kenneth D Forbus, and Ni Lao. 2016. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. *arXiv preprint arXiv:1611.00020* (2016).
- [17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004).
- [18] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).
- [19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
- [20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
- [21] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023* (2016).
- [22] Rodrigo Nogueira, Jannis Bulian, and Massimiliano Ciaramita. 2018. Learning to Coordinate Multiple Reinforcement Learning Agents for Diverse Query Reformulation. *arXiv preprint arXiv:1809.10658* (2018).
- [23] Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-Oriented Query Reformulation with Reinforcement Learning. In *EMNLP*. 574–583.
- [24] Boyuan Pan, Hao Li, Zhou Zhao, Bin Cao, Deng Cai, and Xiaofei He. 2017. MEMEN: multi-layer embedding with memory networks for machine comprehension. *arXiv preprint arXiv:1707.09098* (2017).
- [25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*. ACL, 311–318.
- [26] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.
- [27] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. *ICLR* (2016).
- [28] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International Conference on Machine Learning*. 1889–1897.
- [29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [30] Minjoon Seo, Sewon Min, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Query-reduction networks for question answering. *arXiv preprint arXiv:1606.04582* (2016).
- [31] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.
- [32] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*. 1057–1063.
- [33] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. LSTM-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108* (2015).
- [34] Yi-Lin Tuan, Jinzhi Zhang, Yujia Li, and Hung-yi Lee. 2018. Proximal Policy Optimization and its Dynamic Version for Sequence Generation. *arXiv preprint arXiv:1808.07982* (2018).
- [35] Xin Wang, Wenhui Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160* (2018).
- [36] Ronald J Williams and Jing Peng. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science* 3, 3 (1991), 241–268.
- [37] Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y Ng. 2016. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727* (2016).
- [38] Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 380–386.
- [39] Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *EMNLP*. 584–594.