Multi-turn dialogue-oriented pretrained question generation model

Yanmeng Wang^{1,2} · Wenge Rong^{1,2} · Jianfei Zhang^{1,2} · Shijie Zhou^{1,2} · Zhang Xiong^{1,2}

Received: 17 February 2020 / Accepted: 25 April 2020 / Published online: 16 May 2020
© The Author(s) 2020

Abstract

In recent years, teaching machines to ask meaningful and coherent questions has attracted considerable attention in natural language processing. Question generation has found wide applications in areas such as education (testing knowledge) and chatbots (enhancing interaction). Following previous studies on conversational question generation, we propose a pretrained, encoder–decoder model that can incorporate the semantic information from both passage and hidden conversation representations. We adopt BERT as the encoder to combine external text and dialogue history, and we design a multi-head attention-based decoder to incorporate the semantic information from both text and hidden dialogue representations into the decoding process, thereby generating coherent questions. Experiments with conversational question generation and document-grounded dialogue response generation tasks indicate that the proposed model is superior to baseline models in terms of both standard metrics and human evaluations.

Keywords Multi-turn dialogue · Question generation · Pretrained language model · Text passage

Introduction

Recently, question generation (QG) has attracted considerable interest. Its purpose is to generate human-like questions from a given sentence or paragraph [7]. Owing to its complexity and ambiguity, QG is a highly challenging problem in natural language processing. Specifically, unlike one-to-one mapping tasks such as machine translation, QG involves significant diversity in the space of reasonable questions that may be obtained from a given descriptive text [10].

Current studies on QG use paragraphs and answers as input to predict questions [6,10]. Most QG datasets lack

conversational context, whereas humans normally gather information or test knowledge through conversations involving a series of interconnected questions and answers, thus alleviating ambiguity [28]. To address this issue, Gao et al. recently identified a new QG challenge called conversational QG (CQG) [11]. In this task, a system should be able to generate a series of interconnected questions depending on a given passage, and participate in a question-answering-style conversation as the questioner. Table 1 presents an example of such a task. In this conversation, a questioner and a respondent talk about a passage, and this implies that the content of the passage is important in predicting this conversation. The task here is to evaluate the capability of the system to understand the given passage and output proper and coherent questions corresponding to the provided answers depending on the conversation history. Consequently, a large-scale conversational question-answering (CoQA) dataset [28] has been modified to a CQG dataset by filtering out QA pairs with “yes”, “no”, or “unknown” as answers (28.7% of the total QA pairs). Each sample contains one passage along with conversation history (previous n turns of QA pairs), the current answer, and the question to be predicted.

Gao et al. [11] also presented an end-to-end neural model with coreference alignment and conversation flow modeling. In this model, which outperforms several baseline models, a fundamental requirement is coreference relation annotation

✉ Wenge Rong
w.rong@buaa.edu.cn

Yanmeng Wang
wang.ym@buaa.edu.cn

Jianfei Zhang
zhangjf@buaa.edu.cn

Shijie Zhou
zhoushijie@buaa.edu.cn

Zhang Xiong
xiongz@buaa.edu.cn

¹ State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

² School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Table 1 Example of CQG from CoQA dataset [28]

Passage: As soon as the Columbia could make the proper landing, Captain Ponsberry went ashore and reported his arrival to the authorities, and also reported the escape of Shamhaven and Peterson. The authorities had already heard of the capture of the Columbia from the Russians, and said that the schooner would have to remain at Nagasaki until the whole case could be adjusted. The Japanese were inclined to favor both the Richmond Importing Company and the owners of the vessel, so it was not likely that our friends would lose much in the end. In the meantime, the Columbia could be put in a dry dock and given the overhauling that she needed. “We shall do all we can to locate Shamhaven and Peterson and get back your money”, said an official of the secret service department. But his hands were so full with other matters of greater importance that little attention was paid to the disappearance of the two rascals. “Well, this will tie me up at Nagasaki for some time to come”, said Captain Ponsberry to Larry, on the third day after arriving at the Japanese port.

<i>Q</i> ₁ : Who went ashore?	<i>A</i> ₁ : Captain Ponsberry
<i>Q</i> ₂ : Did he report to anyone?	<i>A</i> ₂ : The authorities
<i>Q</i> ₃ : How many had gotten free?	<i>A</i> ₃ : Two
<i>Q</i> ₄ : Who were they?	<i>A</i> ₄ : Shamhaven and Peterson
<i>Q</i> ₅ : Who captured the Columbia?	<i>A</i> ₅ : Captain Ponsberry
<i>Q</i> ₆ : What type of vessel was it?	<i>A</i> ₆ : Schooner
<i>Q</i> ₇ : What would it have to wait for?	<i>A</i> ₇ : Case could be adjusted
<i>Q</i> ₈ : Where would it be stored?	<i>A</i> ₈ : Dry dock
<i>Q</i> ₉ : What did the authorities think they would get back when they found the men?	<i>A</i> ₉ : Money
<i>Q</i> ₁₀ : When did the captain talk with Larry?	<i>A</i> ₁₀ : Third day

on conversation history as input for training. In the inference phase, the model can generate conversational questions depending on the explicated coreference labels.

Traditional coreference-resolution approaches [9,14] involve complex feature engineering and heavy human labor [33]. Even the best neural coreference-resolution model should compute the scores for all possible spans [21], and this is challenging in dialogue with multiple coreferences [33]. Recently, pretrained language models (e.g., BERT) have been proven effective in coreference resolution [17], as they can capture complex semantic information in long-form language; this is important for multi-turn conversation modeling and the corresponding QG.

In this paper, we propose an advanced CQG framework using text passage fusion and a pretrained language model. We use the BERT model as the encoder, and the decoder in a transformer. The conversation history and relevant text passage are fed into a novel BERT2Trans encoder–decoder model to incorporate the semantic information from both passage and hidden conversation representations.

During fine-tuning, the BERT model should differentiate the input sequences into two segments by a special token “[SEP]” and a learnable segment embedding [3]. However, a long-form text such as a conversation may contain more than two intention or semantic segments [30]; this raises an interesting question: Is it helpful to add more segments into BERT for QG in the context of multi-turn conversation? A conversation consists of multiple turns involving two interlocutors [30], who usually speak alternately. Each turn contains utterances from one speaker. It could be one sentence or multiple sentences. To this end, we further propose an effective multi-task learning mechanism so that segment information of

different turns in the conversation history may be fed into the model during training. Thereby, the proposed BERT2Trans can use input sequences with multiple segments.

To validate the proposed framework, we first test it on the CQG dataset [11]. Furthermore, we also test it on document-grounded conversations (DGCs) [23]. This involves unstructured text passages and multi-turn conversations as input to predict next-turn utterances instead of next-turn questions. Both automatic and human evaluations demonstrate that the proposed model substantially outperforms baseline models and can generate informative and a coherent next-turn utterance (question or dialogue response).

The main contributions of our study are as follows: (1) We propose a novel Bert2trans encoder–decoder using pretrained language models to encode semantic information from both unstructured text and conversation history, and to predict the next-turn question or utterance. (2) We adopt multi-task learning to feed segment information of different turns in a conversation into the model during training. Thereby, the proposed BERT2Trans can use input sequences with multiple segments. (3) The proposed framework is evaluated on two major datasets, and it is demonstrated that it significantly outperforms baseline models and achieves state-of-the-art performance on two generation tasks.

Related work

The purpose of QG is to generate an appropriate question for a given input context and answer. Duan et al. proposed a retrieval-based method using convolutional neural networks and a generation-based method using recurrent neural net-

works (RNNs) [8]. They also used the generated questions to improve existing QA systems and concluded that QG and QA can boost each other.

Seq2Seq models using long short-term memory (LSTM) with global attention are widely used for QG [8,15,18,20,40,42,43]. In decoding, the hidden states of the decoder are used to generate attention weights for the encoded representation of the input. These weights are used to generate the context vector, which is subsequently concatenated with the hidden-state vector for classification. This widely used mechanism may vary in the computation of the attention weights (e.g., Luong or Bahdanau attention).

As the context consists of multiple sentences, Du et al. proposed encoding at sentence and paragraph level [5]. At sentence level, the sum or convolution operation followed by max pooling is applied to the token embedding vector to obtain the representation of each sentence. At paragraph level, the sentence-representation vector is fed into a bi-directional LSTM.

As some of the tokens are rarely used but are also keywords that can appear in both the context and the conversation, a copy mechanism is proposed for directly selecting words from the context [40,42,43]. For repeated words in the context, the copy scores can be naturally higher than those for others. Therefore, Zhao et al. [40] proposed limiting the magnitude of these scores by a maxout pointer. As the question tokens copied from the context tend to be close and relevant to the answer, Zi et al. [43] proposed a position-aware model. Moreover, they proposed selecting question words from a restricted vocabulary. In other models, more input features are added, such as part of speech, named entities, word case, coreference, and dependency [15,20,42,43]. These features are concatenated with the token embedding vector and answer-position signals, and are subsequently fed into an LSTM encoder.

The purpose of CQG is to generate interconnected questions in CoQA [11], where these questions depend on the conversation history and a given passage. CFNet [11] proposed an RNN-based Seq2Seq model with coreference alignment and conversation flow modeling. To obtain a coherent conversation, CFNet adopts a conversation flow mechanism that generates questions about the beginning of a paragraph and gradually shifts focus to the later parts. To correlate the generated question with the dialog history, the coreference alignment mechanism explicitly aligns the referential relationship in the conversation history with the corresponding reference in the generated question.

In parallel to question generation in a conversational context, another closely related line of research is DGC generation, in which dialogue responses are generated based on the context of a given document [23]. The Seq2Seq model has been widely adopted in dialogue response generation by incorporating external unstructured knowledge. External

relevant facts and/or knowledge are another type of information that facilitates dialogue understanding and response generation [12,36,41]. For example, [27] extended Seq2Seq using an external knowledge embedding to incorporate external knowledge from Wikipedia for conversation generation. [4] combined a transformer framework with a memory network to retrieve knowledge and generate natural dialogue responses.

In addition, large-scale unsupervised pretrained language models, such as BERT [3], XLNet [37], and RoBERTa [24] have achieved significant performance gains in several NLP tasks. Some recent studies have also proposed using pretrained language models for text generation [39]. In the present study, we use BERT as an encoder that can fuse information by its multiple self-attention layers in both the right-to-left and left-to-right directions. Furthermore, we adopt multi-task learning so that segment information of different turns in a conversation may be fed into the model during training. Thereby, the proposed BERT2Trans can use input sequences with multiple segments.

Methodology

Overview

The proposed pretrained question-generation framework is shown in Fig. 1. A novel encoder–decoder model uses BERT as the encoder to fuse the information from a given text passage with dialogue history. A multi-head attention-based decoder is proposed to incorporate the semantic information from both the encoded knowledge and hidden dialogue states to generate informative and coherent dialogue responses.

We use $D = \{x_1, \dots, x_m\}$, $P = \{k_1, \dots, k_n\}$, and $Q = \{y_1, \dots, y_l\}$ to represent dialogue history, related text passage, and generated question respectively; x_t , k_t , and y_t are words from a vocabulary V . For a given dialogue history D and the corresponding text passage P , our goal is to generate a proper and informative next-turn utterance Q (question or response in case of DGC).

Encoder

As most question-generation models have an encoder–decoder structure [32,35], we adopt the BERT model [3] to map the input sequence of conversation history and relevant passage into a continuous representation sequence. We propose three different mechanisms to arrange the dialogue history and relevant passage:

- (1) We consider the passage P a sequence of tokens. The conversation history is concatenated with the relevant text

Fig. 1 Knowledge-enhanced response generation

passage, and [SEP] is inserted in between, that is, {[CLS] P [SEP] D [SEP]}.

- (2) In the above method, the multiple-turn utterances of the dialogue history are concatenated into a sequence of tokens in chronological order. Conversation session segmentation (i.e., turn boundaries) is ignored. The model encodes the dialogue history as a continuous sequence of tokens. However, conversation session segmentation is important for semantic understanding. A multi-turn dialogue normally involves two interlocutors [30], who speak alternately. Thus, intentions vary greatly between adjoining turns. In CQG and DGC generation, the last-turn utterance is significantly more relevant to what was generated than those of previous turns. Thus, we propose inserting [SEP] between adjoining turns in the conversation history, that is, {[CLS] P [SEP] Turn₁ [SEP] Turn₂ [SEP] ... [SEP] Turn_N [SEP]}.
- (3) Adopting multiple [SEP] in the input sequence is substantially different from the pretraining setup of BERT [3], and thus the model is difficult to converge during fine-tuning. The self-attention block may pay excessive attention to some tokens in the input sequence. We will elaborate this in detail in Sect. 4.5. We propose multi-task learning so that the segment information of adjacent turns in the conversation may be fed into the model during training. The setup of [SEP] in the input sequence is the same as that in BERT pretraining (i.e., {[CLS] P [SEP] D [SEP]}). As shown in Fig. 1, the boundary detection layer is a simple binary classifier consisting of a linear transformation and a softmax function that map the final hidden vector from BERT to the predicted segmentation labels for each token in the conversation history. The starting tokens of each turn/round are labeled as 1, and other tokens in the turn/round are labeled as 0. In the inference phase, the model can generate the next-turn question without an explicit dialog session segmentation label.

In addition to the position and the token embedding, we also use a learnt segment embedding to every token indicating whether the token belongs to the passage text or dialogue history. As shown in Fig. 1, for each token x_i , the input embedding is the sum of the corresponding token, segment, and position embeddings:

$$I(x_i) = E(x_i) + T(x_i) + P(x_i), \quad (1)$$

where $E(x_i)$, $T(x_i)$, and $P(x_i)$ are the word, segment, and position embeddings, respectively. The positional embedding is a learnable embedding with supported sequence length up to 512 tokens [3].

The input embedding is then fed into the BERT model to obtain encoded representations of the passage and of the dialogue history.

$$H_p; H_d = BERT(I(k_1) \cdots I(k_n), I(x_1) \cdots I(x_m)), \quad (2)$$

where H_p and H_d are the semantic representations of the passage text and the dialogue history, respectively.

Turn-boundary detection output: We assume that H_d^i is the semantic representation of the i th token in the dialogue history. The probability that the i th token is labeled as class c (i.e., the starting token of the next turn) is predicted by the softmax function:

$$P(c|x_i) = \text{softmax}(W_d^T H_d^i), \quad (3)$$

where the parameter W_d is learnt during model training.

Decoder

After comprehensive fusion in the BERT model, H_p and H_d are the semantic representations of the passage text and the dialogue history, respectively. The self-attention sub-layer in

the decoder is modified to prevent paying attention to subsequent positions, ensuring that predictions are only made depending on the known outputs [34]:

$$D_m = \text{MaskedMultiHead}(Q = Q, K = Q, V = Q), \quad (4)$$

where Q is the generated question embedding, and D_m is the output of masked self-attention. Following the definition proposed in [34], Q , K , and V denote the query, key, and value vectors in multi-head attention, respectively.

As shown in Fig. 1, H_d is concatenated with H_p and fused with the response representation in multi-head attention:

$$D_{pd} = \text{MultiHead}(Q = D_m, K = [H_p; H_d], V = [H_p; H_d]). \quad (5)$$

The decoder fuses the information from H_p and H_d , and effectively uses both the conversation context and relevant passage to predict the next-turn utterance. Subsequently, D_{pd} is fed into a position-wise feedforward network.

$$F_{dp} = \text{FFN}(D_{dp}). \quad (6)$$

As in transformer networks [34], the embedding weight matrix of the encoder (BERT model) is shared with the decoder. The pretrained word embedding has rich semantic information about each word [30] and provides a good initialization point to the decoder word embedding.

Copy mechanism

We apply dot-product attention [1] to the decoder hidden states F_{dp} with encoded representations $[H_p; H_d]$ to obtain the attended token representations F'_{dp} and the corresponding attention weight distribution W . The vocabulary distribution M is subsequently calculated using a feedforward neural network:

$$F'_{dp}, W = \text{Attention}(F_{dp}, [H_p; H_d]), \quad (7)$$

$$M = \text{softmax}(\text{FFN}(F'_{dp})). \quad (8)$$

Repeating the words from conversation context is an important ability in human language communication [29]. As some of tokens (e.g., entity names) are rarely used but are also important for context understanding and dialogue generating. The copy mechanism in [13,29] is introduced to allow both copying words from the input by pointing and generating words from a predefined vocabulary V during decoding. It can improve generation accuracy and handle the out-of-vocabulary words in context. In this study, the attended token

representations F'_{dp} and the decoder output F_{dp} are concatenated to learn the generation probability P_{gen} , which is used to obtain the final distribution of the generated question Q :

$$P_{gen} = \text{sigmoid}(\text{FNN}(F'_{dp}; F_{dp})) \quad (9)$$

$$Q = P_{gen}M + (1 - P_{gen})W, \quad (10)$$

where M is the vocabulary distribution from Eq. 8, and W is the attention weight distribution from Eq. 7.

Joint training

We now introduce a turn-boundary detection loss to feed dialogue session segmentation (i.e., boundaries between adjacent turns) information into the model during training:

$$Loss_{boundary} = \frac{1}{N_s} \sum_{n=1}^{N_s} \sum_{i=0}^m \log P(c_i | x_i), \quad (11)$$

where N_s is the number of training samples, m is the number of dialogue history tokens, and c_i is the correct boundary label indicating whether the token x_i is the start of the next turn in the dialogue history.

$$Loss_{gen} = \frac{1}{N_s} \sum_{i=1}^{N_s} \log P(Q_i | D_i, P_i), \quad (12)$$

where N_s is the number of training samples. D_i , Q_i , and P_i are the dialogue history, question, and relevant text passage. $Loss_{gen}$ is the typical negative log-likelihood loss in Seq2Seq learning.

Considering these two components, we define a joint loss function as

$$Loss = \lambda_g Loss_{gen} + (1 - \lambda_g) Loss_{boundary}, \quad (13)$$

where λ_g is hyperparameter.

Experimental study

Datasets

We evaluated the proposed BERT2Trans model on two datasets:

1. CQG dataset. We used a processed version of the dataset in [11]; it contains 66298 training dialog samples and 8360 test samples. A related document from seven different domains [28] is given for each dialogue. The average length of passages, questions, and answers are 332.9, 6.44, and 3.2 tokens, respectively.

2. DGC dataset. We used a processed version of the dataset in [23]. This contains 72922 training samples and 11577 test samples, the utterance style of which is that of a casual chat. A related document about popular movies is given for each dialogue, which contains descriptive information about the movie.

Evaluation metrics and configuration

Regarding the CQG dataset, we used the evaluation metrics BLEU [26] and ROUGE-L [2] to measure the question generation accuracy and make an aligned comparison with CFNet [11]. These metrics are widely used for measuring textual similarity. The distinct-1, distinct-2 [22] metrics were also adopted to measure the degree of diversity of the generated questions. Furthermore, we used the average entropy [31] to measure the performance in providing more informative content in question generation. Regarding the DGC dataset, we closely followed [23] and used perplexity (PPL) and BLEU [26] for comparison. Moreover, we used human evaluation, as it is broadly agreed that objective metrics weakly correlate with human evaluation results. Human evaluation is a necessity in dialogue generation.

Table 2 presents the hyperparameters for training the proposed model. We used the pretrained BERT_{base} as the encoder. The BERT2Trans decoder has two layers of attention blocks. All models were trained using Adam [19] for optimization. The learning rate was set to 1e−4 for the CQG task and 5e−5 for the DGC task. Following [11,23], we used three previous turns of QA pairs in the conversation history for the CQG task, and three utterances for the DGC task to make the aligned comparison. All the models were trained for at most 20 epochs. The best hyperparameter values were selected using the validation set, and the results are reported for the test set.

Experimental results

BERT2Trans+nosep, +sep and, +detect refer to no separation between questions and answers in the conversation history, using “[SEP]” as separation between questions and answers in the conversation history, and BERT2trans with boundary detection mechanism (i.e., the three mechanisms described in Sect. 3.2), respectively.

On the CQG dataset, we compared the proposed models with the following baselines: BahdanauAttention [1] and LuongAttention [25] are Seq2Seq with two different attention mechanisms. PGNet [29] is Seq2Seq with Pointer-generator network. MCT [38] is Seq2Seq with attention, copy mechanism, and answer encoder layer. ER GAN [16] extends attentional Seq2Seq network with an adversarial discriminator and is trained with policy gradient method. NQG [6] is Seq2Seq network that incorporates coreference feature

Table 2 Hyperparameters for model training

Hyperparam		CQG	DGC
Encoder	Number of layers	12	12
	Hidden size	768	768
	Attention heads	12	12
	FFN inner hidden size	3072	3072
Decoder	Number of layers	2	2
	Hidden size	768	768
	Attention heads	8	8
	FFN inner hidden size	2048	2048
Passage max length		460	260
Conversation max length		49	120
Target max length		30	60
Training batch size		48	30
Learning rate		1e−4	5e−5
Beam search size		5	5
Vocab size		30,522	30,522

Table 3 Results for the CQG test set

	B1	B2	B3	B4	R-L
BahdanauAttention	25.52	11.16	6.25	4.09	36.30
LuongAttention	26.89	11.17	5.84	3.58	35.64
PGNet	28.84	13.74	8.16	–	39.18
MCT	35.00	18.24	11.39	7.68	38.18
ER GAN	36.95	19.25	12.07	8.15	39.46
NQG	35.56	21.14	14.84	–	45.58
MSNet	36.27	21.92	15.51	–	46.01
CorefNet	36.89	22.28	15.77	–	46.53
FlowNet	36.87	22.49	15.98	–	46.64
CFNet	37.38	22.81	16.25	12.28	46.90
Bert2trans+nosep	36.45	21.31	14.57	10.62	45.91
Bert2trans+sep	41.99	25.26	17.29	12.49	47.41
Bert2trans+detect	42.89	25.77	17.69	12.69	47.64

Bold values indicate the best results.

via gated network. In BahdanauAttention, LuongAttention, PGNet, MCT, and NQG, passage, conversation history, and current answer are concatenated as a sequence for the input. MSNet [11], CorefNet [11], FlowNet [11], and CFNet [11] are multi-source encoder–decoder models with coreference alignment and conversation flow modeling.

Table 3 shows a comparison of the BLEU (1–4) (i.e., B1, B2, B3, B4) [26] and ROUGE-L (i.e., R-L) [2] scores of the proposed models and those of the baseline models. All the proposed models performed significantly better than the competitive baseline models and achieved state-of-the-art results in CQG [11].

The proposed BERT2Trans+detect outperforms the best baseline (CFNet [11]) by 8.9% and 3.3% in terms of BLEU3

Table 4 PPL and BLEU4 for the DGC test set. Baseline results from [23]

	PPL	BLEU4
Seq2Seq	80.93	0.38
HRED	80.84	0.43
Transformer	87.32	0.36
Wizard Transformer	70.30	0.66
ITE+CKAD	64.97	0.86
ITE+DD	15.11	0.95
BERT2Trans+nosep (ours)	39.89	1.14
BERT2Trans+sep (ours)	38.53	1.19
BERT2Trans+detect (ours)	40.36	1.25
Bert2trans+nosep+DD (ours)	10.52	1.18
Bert2trans+detect+DD(ours)	9.44	1.31

Bold values indicate the best results.

and BLEU4, respectively. As the average question length is only 6.44 tokens, the improvement is distinctly large in terms of BLEU3 and even more significant in terms of BLEU2 and BLEU1. The proposed models achieve better n -gram similarity between the generated question and the ground truth. BERT2Trans+nosep is not better than the baseline model (CFNet). This suggests that separation information of multiple turns (questions and answers) in conversations is important for conversational QG. Good model-generated questions should be based on the answers in the last turn. As shown in Table 3, ER GAN also has room to improve, which is consistent with the finding in previous study [16].

On the DGC dataset, we compared the proposed models with the following baseline models: Seq2Seq [1], HRED [30], Transformer [34], and Wizard Transformer [4]. ITE+CKAD and ITE+DD were proposed in [23] and consist of an incremental transformer with a deliberation decoder for DGCs.

Table 4 shows the PPL and BLEU4 scores for the DGC dataset. The baseline results are reported in [23]. Among the baseline models, ITE+DD has the best performance. This model is enhanced by a deliberation decoder [23, 39]. To make an aligned comparison with [23], we also equipped the proposed Bert2trans+detect model with a similar deliberation decoder. The results demonstrate that the use of the deliberation decoder only improved the PPL score, and the proposed boundary detection mechanism significantly improved BLEU4.

The Bert2trans+detect+DD model outperformed all baseline models in terms of PPL and BLEU4.

Current Seq2Seq models tend to generate generic non-informative text. We present a comparison of the diversity levels of the results for the CQG dataset in Table 5. Li et al. [22] proposed distinct-1 and distinct-2 to measure the diversity of the generated text. These metrics can be computed as the number of distinct unigrams and bigrams divided by

Table 5 Generation diversity results for the CQG test set

	dct-1, dct-2	1gram, 2gram	wd
CFNet	0.072, 0.201	2769, 7735	38k
BERT2Trans+nosep	0.040, 0.101	1457, 3718	37k
BERT2Trans+sep	0.079, 0.245	3388, 10574	43k
BERT2Trans+detect	0.073, 0.245	3390, 11,318	46k

Bold values indicate the best results.

Table 6 Response information entropy for the CQG test set

	U	H_w	H_U
CFNet	4.60	5.78	26.56
BERT2Trans+nosep	4.41	4.79	21.14
BERT2Trans+sep	5.15	6.26	32.26
BERT2Trans+detect	5.52	6.52	35.97
Human	6.44	7.11	45.82

Bold values indicate the best results.

the total number of generated words. Table 5 presents the distinct unigrams, distinct bigrams, and total number of generated words as 1gram, 2gram, and |wd|, respectively. The proposed BERT2Trans+detect model substantially increased dct-1 and dct-2 (i.e., the distinct unigrams and bigrams, respectively) over the baseline model (CFNet [11]). This indicates that the proposed models can significantly improve the diversity of the generated questions.

The BERT2Trans+sep model achieved the best distinct-1 score. It yielded smaller values for 1gram, 2gram, and |wd|. A high distinct-1 score can be a direct consequence of the relatively low total number of generated words. As shown in Table 5, the proposed models generated more diverse questions compared with the baseline model (CFNet [11]).

Table 6 shows the average response length and average response information entropy with respect to the maximum likelihood unigram model for the generated questions. Following the definition in [31], we computed the unigram probabilities based on the maximum-likelihood unigram distribution of the training corpus. H_w (i.e., information entropy per word) is computed as [31]

$$H_w = - \sum_{w \in U} p(w) \log p(w), \quad (14)$$

H_U denotes the information entropy per response and |U| denotes the average response length. It can be observed that the proposed models, particularly BERT2Trans+detect, generated questions with larger average length and remarkably enhanced the utterance entropy H_U . This indicates their capability to generate questions with higher informativeness, and it is in good agreement with the previous experiment regarding the BLEU and ROUGE-L metrics. The human-generated questions were superior to all questions generated by neural

Table 7 QG examples**Case 1:****Passage:**

...just pop pin in to say I let my balloon off with a message on it, hope you got it ok and it made you laugh up there”. Five more men aged between 17 and 27 have been found hanged in the area since January 2007. Speaking to the Daily Mail newspaper, Liam Clarke’s father, Kevin Clarke, said the seven who had killed themselves appeared to have known each other. “we don’t know if it is some weird cult or copycat suicides or if they have had some bizarre pact to kill themselves”, Clarke said.

Conversation history: Q: Who is Liam Clarke’s father? A: Kevin Clarke, Q: Who did he speak to? A: The newspaper, Q: Did the people who did this know each other? A: Yes

Answer: Daily Mail

Gold: Which newspaper did he talk to?

CFNet (baseline): How often?

BERT2Trans+detect (proposed): Which newspaper?

Case 2:**Passage:**

...Kenny said he developed his love of cooking by watching his mother, and his grandmother, who owned a catering business herself. Kenny helped them both in order to remember their tips: How long to cook chicken so it stays wet, and the right amount of tomatoes to add to a spaghetti dish. At the age of 5, he cooked his first dish of shrimp and broccoli ...

Conversation History: Q: How did he obtain his passion for cooking? A: Watching his mother and grandmother, Q: Did he learn from them? A: Yes, Q: What tips he did get? A: How long to cook chicken, how many tomatoes go in a spaghetti dish

Answer: At the age of 5

Gold: How old was he when he started cooking?

CFNet (baseline): How old was he?

**BERT2Trans+detect
(proposed):**

When did he get
his first dish of
cooking?

Case 3:**Passage:**

...Kournikova made her WTA first show at 15 years old at the US Open where she finally lost against player Steffi Graf. But she made it to the double quarter finals that same match. In 1996, Kournikova won the rookie of the year award and the next year ... aim in 1999, she made her first career WTA final in Key Biscayne against Venus Williams in a tough 3 set match. She also won her first doubles title with Monica Seles in Tokyo. At present Kournikova is more successful on the net than at the net. She remains the “most searched” and “most download” on the internet, three times more popular than the no. 2 sports figure, Michael Jordan. She is still very young and she seems to have a great future ahead!

Conversation history: Q: Where is she most successful? A: Doubles, Q: More popular than whom? A: Michael Jordan, Q: In 1999, she battled whom in a tough match? A: Venus Williams

Answer: 1996

Gold: When was she rookie of the year?

CFNet (baseline): When?

**BERT2Trans+detect
(proposed):**

What year did she
win the rookie of
the year?

Table 8 Examples of generated dialog responses for DGC**Case 1:****movie passage:**

The Wolf of Wall Street is a 2013 American biographical black comedy crime film directed by Martin Scorsese and written by Terence Winter, based on the memoir of the same name by Jordan Belfort. It recounts Belfort's perspective on his career as a stockbroker in New York City and how his firm, Stratton Oakmont, engaged in rampant corruption and fraud on Wall Street, which ultimately led to his downfall. Leonardo Dicaprio (who was also a producer) stars as Belfort, with Jonah Hill as his business partner and friend Donnie Azoff, Margot Robbie as his wife Naomi Lapaglia and Kyle Chandler as Patrick Denham, the FBI agent who tries to bring him down.

Dialog history:

T1: Hello. Who is main character in your movie?

T2: Leonardo Dicaprio

T3: I love him! what movie is it?

Responses:

Gold: Wolf of Wall Street

ITE+DD (base-line): It is based on memoir of Jordan Belfort

Bert2Trans+detect+DD (proposed): The Wolf of Wall Street

Case 2:**movie passage:**

The Duo try to make it to the moving truck, but Sid's dog Scud sees them and gives chase. Buzz saves Woody from Scud but is left behind, so Woody attempts to rescue him with Andy's remote-controlled car, RC. Thinking that Woody is "killing" RC as well, the other toys attack and toss him off the truck. Having evaded Scud, Buzz and RC pick up Woody and continue after the truck. Upon seeing Woody and Buzz together on RC, the other toys realize their mistake, and try to help them get back aboard, but RC's batteries become depleted, stranding them. ... Woody and Buzz stage another reconnaissance mission to prepare for the new toy arrivals...

Dialog history:

T1: Wow I guess Buzz impresses not only toys but humans. Toys can be vicious.

T2: Yes, they can be. However, there is an upside. Buzz and Woody become a duo as they fight off things like Sid and his dog.

T3: Sid and his dog are supposed to be evil. So the toys learn to work together?

Responses:

Gold: Sid is mischievous. And yes toys can learn to work together.

ITE+DD (baseline): I don't know.

Bert2Trans+detect+DD (proposed): Yes, they learn to work together.

Fig. 2 Visualization of attention weights in the last self-attention layer of BERT encoder

generative models in terms of response length and information entropy, suggesting that a higher entropy is desirable [31].

Case study

Table 7 shows three QG examples from the CQG task, where the proposed BERT2Trans+detect model substantially outperformed the baseline model. In case 1, it appears that the baseline model failed to learn the semantics of interconnected questions in the conversation history and generated a question related to the word “daily”, whereas our model generated a coherent question depending on the conversation history and the passage. In case 2, although the answer is “at the age of 5”, the passage and conversation history are about “Kenny’s love of cooking”. The proposed model determined the sentences related to the conversation topic and generated a question related to the time when he cooked his first dish. In case 3, the baseline model only generated a generic question “when?” and completely ignored the story of the tennis ball player. The proposed model generated a specific question using the correct contextual information.

Table 8 shows the generated responses for the DGC dataset by the proposed Bert2Trans+detect and ITE+DD [23], which achieves the best performance in terms of automatic evaluation. In case 1, both Bert2Trans+detect and ITE+DD output meaningful responses depending on the relevant movie descriptions; however, the Bert2Trans+detect model output a more proper response in the context of conversation. In case 2, ITE+DD generated a generic response, whereas the Bert2Trans+kg model produced the correct answer to the question. This indicates that the proposed model understands implicit semantic information in long-form language.

Attention visualization

Figure 2 shows a comparison of the attention weights in the last self-attention layer of the BERT model (encoder) for three mechanisms for separating questions and answers (or turns) in the conversation history (see Sect. 3.2). As shown in Fig. 2a, the setup of the input sequence is same as that for pretraining (i.e., $\{[CLS] K [SEP] D [SEP]\}$). All words pay considerable attention to “[SEP]”. Figure 2b shows that if multiple “[SEP]s” are inserted to separate questions and answers in the conversation history, the model will pay great

Table 9 Human evaluation for the CQG dataset

	CFNet	BERT2Trans+detect
Grammaticality	2.61	2.67
Answerability	2.14	2.35
Interconnectedness	2.29	2.37

attention to other tokens (i.e., “.”). The additional “[SEP]” misleads the BERT model to learn an abnormal attention weight. By contrast, “[SEP]” does not receive significant attention. Figure 2c, d shows the weight of two self-attention heads. As shown on the *x*-axis, the conversation history is “Q: who knocked on the door? A: the lawyer Q: when? A: at eight o’clock Q: who did he find? A: the artist A: sadly altered for the worse”. Without explicit indication in the test data, the proposed model learns the correct boundary between adjacent questions and answers, as demonstrated by the large attention weight received by “[SEP]” (Fig. 2a). The last answer (i.e., “sadly altered for the worse”) is the most important information for generating an informative and coherent question. As shown in Fig. 2d, one of the self-attention heads pays special attention to the starting token of the current answer; this is not found in the previous two mechanisms (see Sect. 3.2).

Human evaluation

We also conducted a manual evaluation for both the CQG and DGC datasets, as a supplement to the evaluation through the metrics. For the CQG dataset, 93 randomly sampled questions with relevant passage, conversation history, and current answer were used for human evaluation. Human annotators evaluated the questions generated by the baseline model (CFNet) and those generated by the proposed model in terms of the same metrics as in [11]: “Grammaticality”, “Answerability”, and “Interconnectedness”. “Grammaticality” measures whether the generated question is fluent and grammatically correct. “Answerability” measures whether the generated question can be answered by the current answer [11]. “Interconnectedness” evaluates whether the generated question can refer back to the conversation history [11].

For the DGC dataset, we randomly selected 100 samples for human evaluation. We used the same metrics as in [23] (i.e., “Fluency”, “Knowledge Relevance”, and “Context Coherence”) to compare the proposed model with the best baseline model (ITE-DD) [23]. “Fluency” measures the fluency of a response, “Knowledge Relevance” determines whether relevant knowledge was used in a response, and “Context Coherence” considers whether the responses are coherent with dialogue history.

Table 10 Human evaluation for DGC dataset

	ITE+DD	BERT2Trans+detect+DD
Fluency	2.53	2.55
Knowledge Relevance	1.34	1.87
Context Coherence	1.87	2.19

We anonymized the model identities for each generated response. All metrics are on a 1–3 scale (3 for the best). The results are shown in Tables 9 and 10. The proposed models outperformed the baseline models (CFNet and ITE+DD) in terms of all six metrics. Both the proposed and the baseline model achieved high “Grammaticality” and “Fluency” scores; this is consistent with the findings in [11, 23]. It is conceivable that both the baseline and the proposed model can generate fluent and grammatical next-turn utterances (question or response). In addition, the proposed BERT2Trans + detect model is superior to the baseline models in terms of the “Answerability” and “Knowledge Relevance” metrics, suggesting that BERT2Trans can better combine passages and dialogue history, and can generate appropriate next-turn utterances (question and response) with better knowledge relevance.

Conclusion and future work

In this paper, we proposed a deep pretrained question generation model to facilitate conversation history understanding and question generation by using the information fusion ability of pretrained language models. We evaluated the model on CQG and DGC datasets in terms of multiple metrics. The results demonstrated that the proposed approach can generate more coherent and informative next-turn utterances (question or response).

In future work, we should enhance the model with reasoning ability for multiple passages or knowledge sources. Thereby, the most relevant textual knowledge content may be distinguished, and the effectiveness of conversation generation may be improved. The experiments and the observations regarding the generated utterances indicate that the comprehensive representation of the complexity of multi-turn conversation remains challenging for current models.

Acknowledgements This work was partially supported by the State Key Laboratory of Software Development Environment of China (No. SKLSDE-2019ZX-16).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material

in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations
- Chin-Yew L (2014) ROUGE: A package for automatic evaluation of summaries. In: Proceedings of 2014 Workshop on Text Summarization Branches Out
- Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 4171–4186
- Dinan E, Roller S, Shuster K, Fan A, Auli M, Weston J (2019) Wizard of wikipedia: Knowledge-powered conversational agents. In: Proceedings of the 7th International Conference on Learning Representations
- Du X, Cardie, C (2017) Identifying where to focus in reading comprehension for neural question generation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp 2067–2073
- Du X, Cardie, C (2018) Harvesting paragraph-level question-answer pairs from wikipedia. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp 1907–1917
- Du X, Shao J, Cardie C (2017) Learning to ask: Neural question generation for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp 1342–1352
- Duan N, Tang D, Chen P, Zhou M (2017) Question generation for question answering. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 866–874
- Durrett G, Klein D (2013) Easy victories and uphill battles in coreference resolution. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp 1971–1982
- Gao Y, Bing L, Chen W, Lyu MR, King I (2019) Difficulty controllable generation of reading comprehension questions. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp 4968–4974
- Gao Y, Li P, King I, Lyu MR (2019) Interconnected question generation with coreference alignment and conversation flow modeling. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, pp 4853–4862
- Ghazvininejad M, Brockett C, Chang M, Dolan B, Gao J, Yih W, Galley M (2018) A knowledge-grounded neural conversation model. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp 5110–5117
- Gu J, Lu Z, Li H, Li VOK (2016) Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 1631–1640
- Haghighi A, Klein D (2009) Simple coreference resolution with rich syntactic and semantic features. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1152–1161
- Harrison V, Walker MA (2018) Neural generation of diverse questions using answer focus, contextual and linguistic features. In: Proceedings of the 11th International Conference on Natural Language Generation. pp. 296–306
- Hosking T, Riedel S (2019) Evaluating rewards for question generation models. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2278–2283
- Joshi M, Levy O, Zettlemoyer L, Weld DS (2019) BERT for coreference resolution: Baselines and analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 5802–5807
- Kim Y, Lee H, Shin J, Jung K (2019) Improving neural question generation using answer separation. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. pp. 6602–6609
- Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations
- Kumar V, Boorla K, Meena Y, Ramakrishnan G, Li Y (2018) Automating reading comprehension by generating question and answer pairs. In: Proceedings of 22nd Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. pp. 335–348
- Lee K, He L, Lewis M, Zettlemoyer L (2017) End-to-end neural coreference resolution. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 188–197
- Li J, Galley M, Brockett C, Gao J, Dolan B (2016) A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 110–119
- Li Z, Niu C, Meng F, Feng Y, Li Q, Zhou J (2019) Incremental transformer with deliberation decoder for document grounded conversations. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. pp. 12–21
- Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach. CoRR, vol. abs/1907.11692
- Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1412–1421
- Papineni K, Roukos S, Ward T, Zhu W (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318
- Parthasarathi P, Pineau J (2018) Extending neural generative conversational model using external knowledge sources. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 690–695
- Reddy S, Chen D, Manning CD (2019) CoQA: A conversational question answering challenge. Trans Assoc Comput Linguist 7:249–266
- See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 1073–1083
- Serban IV, Sordani A, Bengio Y, Courville AC, Pineau J (2016) Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. pp. 3776–3784
- Serban IV, Sordani A, Lowe R, Charlin L, Pineau J, Courville AC, Bengio Y (2017) A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. pp. 3295–3301

32. Shang L, Lu Z, Li H (2015) Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. pp. 1577–1586
33. Su H, Shen X, Zhang R, Sun F, Hu P, Niu C, Zhou J (2019) Improving multi-turn dialogue modelling with utterance rewriter. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. pp. 22–31
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 2017 Annual Conference on Neural Information Processing Systems. pp. 5998–6008
35. Vinyals O, Le QV (2015) A neural conversational model. CoRR, vol. abs/1506.05869
36. Wang Y, Rong W, Ouyang Y, Xiong Z (2019) Augmenting dialogue response generation with unstructured textual knowledge. IEEE Access 7:954–963
37. Yang Z, Dai Z, Yang Y, Carbonell JG, Salakhutdinov R, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. In: Proceedings of the 2019 Annual Conference on Neural Information Processing Systems. pp. 5754–5764
38. Yuan X, Wang T, Gülçehre Ç, Sordoni A, Bachman P, Zhang S, Subramanian S, Trischler A (2017) Machine comprehension by text-to-text neural question generation. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. pp. 15–25
39. Zhang H, Cai J, Xu J, Wang J (2019) Pretraining-based natural language generation for text summarization. In: Proceedings of the 23rd Conference on Computational Natural Language Learning. pp. 789–797
40. Zhao Y, Ni X, Ding Y, Ke Q (2018) Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3901–3910
41. Zhou H, Young T, Huang M, Zhao H, Xu J, Zhu X (2018) Commonsense knowledge aware conversation generation with graph attention. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 4623–4629
42. Zhou Q, Yang N, Wei F, Tan C, Bao H, Zhou M (2017) Neural question generation from text: A preliminary study. In: Proceedings of the 6th CCF International Conference on Natural Language Processing and Chinese Computing. pp. 662–671
43. Zi K, Sun X, Cao Y, Wang S, Feng X, Ma Z, Cao C (2019) Answer-focused and position-aware neural network for transfer learning in question generation. In: Proceedings of the 12th International Conference on Knowledge Science, Engineering and Management. pp. 339–352

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.