Towards Answer-unaware Conversational Question Generation

Mao Nakanishi  Tetsunori Kobayashi  Yoshihiko Hayashi
School of Science and Engineering, Waseda University
Waseda-machi 27, Shinjuku, Tokyo 1690042, Japan
nakanishi@pci.cs.waseda.ac.jp  koba@waseda.jp  yshk.hayashi@aoni.waseda.jp

Abstract

Conversational question generation is a novel area of NLP research which has a range of potential applications. This paper is first to present a framework for conversational question generation that is unaware of the corresponding answers. To properly generate a question coherent to the grounding text and the current conversation history, the proposed framework first locates the focus of a question in the text passage, and then identifies the question pattern that leads the sequential generation of the words in a question. The experiments using the CoQA dataset demonstrate that the quality of generated questions greatly improves if the question foci and the question patterns are correctly identified. In addition, it was shown that the question foci, even estimated with a reasonable accuracy, could contribute to the quality improvement. These results established that our research direction may be promising, but at the same time revealed that the identification of question patterns is a challenging issue, and it has to be largely refined to achieve a better quality in the end-to-end automatic question generation.

1 Introduction

Research on question generation has attracted considerable attention from NLP community, and several neural network-based methods have been proposed (Pan et al., 2019). Many of these methods are developed for text-based question answering (QA) with stand-alone interactions. That is, QA pairs is basically independent each other. Besides, they are generally answer-aware: a question generation system presumes that the corresponding answer to a to-be-generated question is being supplied.

One of the recently emerging directions in QA is conversational QA, in which a series of interrelated QA turns is performed. Within this trend, Gao et al. (2019) recently proposed a framework for conversational question generation. The proposed work is reported effective, but still answer-aware, which may prevent the proposed framework to be applied to practical applications such as chatbots and dialogue systems: answers are usually not provided in the usage scenarios.

Being motivated by this situation, the present work is first to propose a framework for answer-unaware conversation question generation, by assuming that questions coherent to the target text and the current conversation history can be generated, provided the question focus and the question type are properly identified. To confirm this assumption, we have developed a deep neural architecture for answer-unaware question generation, which first tries to locate the focus of a question in the grounding text passage, and then identify the question type that leads the sequential generation of the words in a question.

The experiments using the CoQA dataset (Reddy et al., 2019) demonstrate that the quality of generated questions greatly improves if the question foci and the question patterns are correctly identified. Besides, it was shown that the question foci can be estimated with a certain degree of accuracy, and the quality of the generated questions referring the question foci are superior to that generated from the whole text passage, suggesting that the proper narrowing down of the source of question is essential. These results established that our research direction may be promising. However, it was also proved that it difficult to correctly estimate the question pattern, and the wrongly-identified question patterns severely affect the quality of generated questions. This result may highlight the necessity of incorporating additional clues, such as entities in the text, and developing a refined model to better consume the enriched input information.
2 Related Work

Given a range of application areas, such as intelligent tutoring systems, dialogue systems and question answering systems, question generation has attracted larger research attention in NLP community. The major trend in question generation has shifted from template-based generation systems to neural network-based end-to-end methods (Pan et al., 2019), which generally employs encoder-decoder models. Succeeding the pioneering work (Du et al., 2017), several proposals (Zhou et al., 2017; Tang et al., 2017) have been made to chiefly improve the quality of generated questions. These methods all deal with text-based question answering (MRQA) research. In the context of the present work, however, it should be noted that answering (MRQA) research. In the context of the present work, however, it should be noted that the majority of these methods are answer-aware, which may somehow restrict its application areas, in particular such as dialogue systems. Thus answer-unaware conversational question generation first to offered by the present work would be a natural research direction to go.

3 Framework for Conversational Question Generation

3.1 Overview

Figure 1 overviews our proposed framework for CQG, where the following assumptions are made.

- A question coherent to the current conversational context can be generated primarily by knowing the current focus of interrogation, even without knowing the pre-defined corresponding answer. We herein expect that a question focus can be properly estimated as a textual region in the given passage by exploiting conversation history.

- The quality of a question can be further improved, if the type of a question is identified ahead of time. We consider that the question pattern that linguistically realizes a question type could be identified by using the estimated question focus.

3.2 Problem Formulation

The generation of a conversational question $Q_i$ at the current ($i$-th) QA turn is formulated as follows.

$$Q_i = \arg \max_{Q_i} \text{Prob}(Q_i | P, H_i)$$

Here, $P$ denotes the whole text passage provided for the QA session, and $H_i$ dictates the current conversation history, which can be formulated as $H_i = ((Q_1, A_1), \ldots, (Q_{i-1}, A_{i-1}))$. Notice that the answer $A_i$ corresponding to the to-be-generated question $Q_i$ is not included in our problem formulation.

Question Focus Estimation: We assume that a question focus $F_i$ can be located at a textual region in the grounding text passage $P$, meaning that the answer of a to-be-generated question can be found in this textual region. Given the conversation history $H_i$, the estimation of a question focus is formulated as a classification problem which identifies the most probable text chunk $P_i$ from the $N_C$-divided passages $P = (P_1, \ldots, P_{N_C})$.

Question Pattern Identification: We expect by additionally knowing the type of a question, such as When, Who, Where, and Did, the quality of a generated question may further improve. As detailed in the next section, we cast the identification of a question type as the classification from an inventory of question patterns, or as the actual generation of a question-leading linguistic expression. As discussed in the later section, we experimentally compare these two methods. We denote a question pattern $T_i$ as an element defined in the set of question patterns $T_Q = \{T_1, \ldots, T_{N_T}\}$. $T_Q$ has been mined, in the present work, from the target dataset.
**Question Decoding:** The conversational question generation as formulated in Eq.1 can be further conditioned by incorporating the estimated question focus $F_i$, and the identified question pattern $T_i$. We employ a conventional encoder-decoder model for this process.

\[
\tilde{Q}_i = \arg \max_{Q_i} \text{Prob}(Q_i | P, H_i, F_i, T_i) \quad (2)
\]

### 4 Model Description

This section details the components in the proposed framework, which are (1) Question focus estimation, (2) Question pattern identification, and (3) Question decoding.

Let us assume that the current time step is $t = i$ in the following descriptions. The input to the entire question generation system is the target text passage $P$ and the current conversation history $H_i$.

The passage $P$ is segmented into a sequence of $N_c$ chunks ($P_1, \ldots, P_{N_c}$), where the $c$-th chunk $P_c = (w_{p_1}^c, \ldots, w_{p_m}^c)$ is a sequence of $m$ word tokens.

Although the conversation history $H_i$ at the $i$-th QA turn is conceptually defined as $H_i = ((Q_1, A_1), \ldots, (Q_{i-1}, A_{i-1}))$, we implement it as the sequence of words taken from the question and the answer, separated by a separator: $H_i = (\ldots, w_{q_{i-1}}^t \ldots w_{q_1}^t, \langle \text{sep} \rangle, w_{a_1}^t, \ldots, w_{a_{i-1}}^t \ldots \ldots)$.

We henceforth abbreviated it as $H_i = (w_{H_1}^t \ldots w_{H_{i-1}}^t)$.

The question focus $F_i$ for the $i$-th QA turn is estimated as one of the chunks. It is hence denoted as a sequence of $m$-word tokens: $F_i = (w_{F_1}^t \ldots w_{F_m}^t)$.

The question pattern $T_i$ that is identified for a to-be-generated question is chosen from the pre-defined set $T_Q$ of linguistic expressions, or generated on-the-fly. It is formulated as a sequence of $l$ word tokens: $T_i = (w_1^{T_i}, \ldots, w_l^{T_i})$.

### 4.1 Question Focus Estimation

Figure 2 models the deep architecture for estimating a question focus, which consists of embedding layer, contextual layers, attention layer, modeling layer, and output layer.
dimensional embedding vector for the $i$-th word token in $E^{pc}$. We employ GloVe (Pennington et al., 2014) vectors ($d = 300$) as word embeddings. Similarly we map a conversation history $H_i$ to $E^{H_i} = (e_1^{H_i}, \ldots, e_n^{H_i}) \in \mathbb{R}^{n \times d}$.

Two contextual layers, one is for passage chunks and the other is for conversation history, are both implemented by using Bi-GRU. The input to the passage context layer for a chunk $P_c$ is the concatenation of $E^{pc}$ and $f^{QF}_{i-1}$. The latter vector $f^{QF}_{i-1}$ carries important information in the sense that it specifies the question focus at the previous time step ($t = i - 1$). The elements of $f^{QF}_{i-1}$ are all one if $F_{i-1} = P_c$, otherwise they are all zero. The representation of the current conversation history $E^{H_i}$ is also fed into the contextual layer. The resulting contextual representations $H^{Pc} \in \mathbb{R}^{m \times 2v}$ and $H^{Hi} \in \mathbb{R}^{n \times 2v}$ are fed into the attention layer. Here $v$ represents the dimensionality of the hidden layers: $v = 128$ in our experiments.

The attention layer captures the relative importance of each chunk seeing from the current conversation history as an attentional weight, and hence yields history-augmented contextual representations for the chunks, as formulated below. Here, $W_e$ and $W_h$ are trainable parameters.

\begin{align}
    e^f_{t,j} &= \tanh(W^f_e[h^c_i; h^H_j]) \quad (3) \\
    \alpha^f_{t,j} &= \frac{\exp(e^f_{t,j})}{\sum_{k=1}^{n} \exp(e^f_{t,k})} \quad (4) \\
    c^f_i &= \sum_j \alpha^f_{t,j} h^c_j \quad (5) \\
    \tilde{h}^c_i &= \tanh(W^f_h[c^f_i; h^H_i]) \quad (6)
\end{align}

The modeling layer is also realized by employing Bi-GRU, which captures interactions among the history-augmented contextual representations. That is, we expect that the resulting representation for a chunk $M^{c_i} \in \mathbb{R}^{m \times 2v}$ incorporates relevant information from the conversation history.

The output layer, consists of two linear layers, predicts the most probable chunk index $y_{Fi}$, which means that the designated chunk is estimated as the current question focus $F_i$. The inputs to this layer is $[M^{c_1}; M^{c_2}; \ldots; M^{c_N}] \in \mathbb{R}^{(N \times m) \times 2v}$, which is the concatenation of the chunk representations yielded by the modeling layer.

### 4.2 Question Pattern Identification

The proper identification of a question pattern help improve the quality of a generated question. We approach this task by either of classification or generation, and experimentally compare them.

#### 4.2.1 Question Pattern Classification

As displayed in Figure 3, the whole structure of the classification model is similar to that of the question focus estimation model. This model however only considers the chunk that is estimated as the current question focus. More specifically, the question focus is represented as $[E^{Fi}; f^{NE}_{Fi}]$. That is, the original representation for question focus $E^{Fi}$ is enhanced by the named-entity (NE) tag features $f^{NE}_{Fi} \in \mathbb{R}^{m \times 18}$. We assign to each word token in $F_i$ an NE tag with the BIO format. We use spaCy\(^1\) as the NE recognizer, which maintains 18 NE types\(^2\).

![Figure 3: Question pattern classification model.](https://spacy.io)

The history-augmented representation of the question focus $H^{Fi}$, yielded by the attention and

---

\(^1\)https://spacy.io

\(^2\)https://spacy.io/api/annotation#named-entities.
the modeling layers, is then fed into the output layer, and the index of the most probable question pattern $y_{T_i} \in \mathbb{R}^{NP}$ is finally obtained, where $NP$ represents the number of pre-defined question patterns.

4.2.2 Question Pattern Generation

As illustrated in Figure 4, the generation model only differs from the classification model at the output layer: instead of the classification layer, this model naturally employs a conventional encoder-decoder layers for generating a question pattern.

The encoder takes the question focus $\vec{H}_F$ as the input, and encodes its word token sequence by employing Bi-GRU. The decoder generates the most probable question pattern $P_i$ as a sequence of word tokens $(w_1^P, \ldots, w_l^P)$, while attending to relevant parts in the question focus chunk.

\[
\begin{align*}
    s_t &= \text{GRU}(w_{t-1}^P, c_{t-1}, s_{t-1}) \\
    e_{t,j} &= \tanh(W_q^e s_{t-1} + U_q^e h_E^{t-1}) \\
    \alpha_{t,j}^q &= \frac{\exp(e_{t,j}^q)}{\sum_{k=1}^{n} \exp(e_{t,k}^q)} \\
    c_t^q &= \sum_j \alpha_{t,j}^q h_E^q \\
    \tilde{h}_t^q &= \tanh(W_h^q[c_t^q; h_E]) \\
    p(w_t^P | w_{<t}^P, h_i) &= \text{softmax}(W_d h_i^q)
\end{align*}
\]

4.3 Question Decoding

The question decoding model also employs a conventional encoder-decoder model with attention. Its behavior depends on whether a predicted/generated question pattern is employed. That is, when a question pattern is not used, the input to the encoder is only the representations for a question focus $F_i$. On the other hand, in the latter case, the input to the encoder is the concatenation of the representation for the predicted/generated question pattern $T_i = (w_1^P, \ldots, w_l^P)$ and the question focus chunk $F_i = (w_1^F, \ldots, w_m^F)$, delimited by the separator $(sep)$.  

5 Experiments

5.1 Dataset

The present work relies on the CoQA dataset (Reddy et al., 2019) in the evaluation as well as the model training, which enables us to compare our results with the most relevant related
Figure 6: Example of a QA conversation in CoQA; adopted from (Reddy et al., 2019).

Q1: What are the candidates running for?
R1: The Virginia governor’s race

Q2: Where?
A2: Virginia
R2: The Virginia governor’s race

Q3: Who is the democratic candidate?
A3: Terry McAuliffe
R3: Democrat Terry McAuliffe

Q4: Who is his opponent?
A4: Ken Cuccinelli
R4: Republican Ken Cuccinelli

Q5: What party does he belong to?
A5: Republican
R5: Republican Ken Cuccinelli

Q6: Which of them is winning?
A6: Terry McAuliffe
R6: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn’t trailed in a poll since May

Table 1 displays some question patterns and their frequencies. We train the question pattern identification models by limiting the number of examples to at most 300 to avoid the data imbalance across the patterns.

| Pattern | Raw count | Frequency (%) |
|---------|-----------|---------------|
| what    | 32098     | 29.5          |
| who     | 15692     | 14.4          |
| ...     | ...       | ...           |
| what did| 5636      | 5.19          |
| what did he| 1801 | 1.66 |
| ...     | ...       | ...           |
| UNKOWN  | 2898      | 2.67          |

Table 1: Question patterns ($n = \{1, 2, 3\}, N = 200$).

Comparing baselines: Two baseline question generation systems are employed.

- NQG (Du et al., 2017) is used to assess the efficacy of question focus prediction and question pattern identification. We consider the whole passage as a single chunk when using this system. This means that a question focus is not narrowed down to some textual region, rather it spreads to the whole passage.

- CFNet (Gao et al., 2019), the only known CQG system, is adopted to chiefly evaluate the impact of answer-unawareness. This system still requires the corresponding answer to be supplied to generate a question, although it may be superior to our system in that it is equipped with explicit mechanisms to deal with coreference and conversation flow.
6 Results and Discussions

6.1 Quality of the Generated Questions

The results shown in Table 2 establish our primary assumption, which states that a question coherent to the current conversational context can be generated primarily by knowing the current focus of interrogation. As shown in the table, the qualities of generated questions (as measured by BLEU 1-4), when a question focus is estimated ($N_c > 1$), were better than that from the case where the whole text passage was simply considered as a question focus ($N_c = 1$). These results indeed dictate that the notion of question focus is effective.

| $N_c$ | B1 | B2 | B3 | B4 |
|-------|----|----|----|----|
| 1 (whole passage) | 30.19 | 12.85 | 0.32 | 0.13 |
| 5 (random) | 33.83 | 16.08 | 0.59 | 0.13 |
| 5 (predicted) | 34.64 | 16.65 | 0.70 | 0.18 |
| 10 (predicted) | 34.71 | 16.68 | 0.70 | 0.17 |
| 5 (GT) | 34.19 | 16.30 | 0.71 | 0.21 |
| 10 (GT) | 34.71 | 16.67 | 0.73 | 0.21 |

Table 2: Qualities (BLEU scores) of generated questions (without considering question patterns).

The table further shows that the qualities of generated questions were slightly better than that from the random choice of a chunk as question focus, suggesting that the incorporation of even an estimated question focus is effective. The displayed results, on the other hand, shows that the quality of generated questions (B1 around 34.6) is still not suffice by only knowing the question foci, suggesting the necessity of additional information.

Given these discussions, Table 3 displays the qualities of generated questions under several conditions, and it confirms the above mentioned prospect may be probable. The major outcomes provided in the table are: (1) the generation quality could be largely improved if the focus and the pattern of the to-be-generated question are correctly identified, and (2) the current question pattern identification models severely suffer from the low accuracies, even with classification or generation, and they are comparable or only slightly better than the Random baseline, largely affecting the final generation results.

Table 3: Qualities (BLEU scores) of generated questions.

| $N_c$ | Focus | Pattern | B1 | B2 | B3 | B4 |
|-------|-------|---------|----|----|----|----|
| 5 | P | Gen | 24.15 | 9.80 | 0.14 | 0.02 |
| 5 | P | Class | 27.62 | 13.67 | 0.13 | 0.04 |
| 5 | P | Random | 27.35 | 13.70 | 0.17 | 0.03 |
| 10 | P | Gen | 32.36 | 16.06 | 0.37 | 0.04 |
| 10 | P | Class | 26.87 | 13.00 | 0.16 | 0.04 |
| 10 | P | Random | 28.45 | 14.43 | 0.20 | 0.04 |
| 5 | GT | GT | 56.22 | 38.84 | 18.69 | 7.10 |
| 10 | GT | GT | 53.05 | 34.17 | 14.23 | 7.10 |

Table 4: Comparison of the qualities (BLEU scores) with the baseline systems: NQG (Zhou et al., 2017) and CFNet (Gao et al., 2019).

Table 5 measures the accuracy of query focus estimation with varying $N_c$. The accuracy figures presented in the table may be reasonable, if not satisfactory. The longer chunks achieve apparently higher classification accuracies, but there may be a trade-off between the quality of generated questions. A bigger textual region may not well constrain the content of a to-be-generated question.

| $N_c$ | Ave. Chunk Length | Accuracy (%) |
|-------|-------------------|--------------|
| 5     | 120               | 59.78        |
| 10    | 60                | 48.17        |

Table 5: Accuracy of question focus estimation.

6.2 Accuracy of Question Focus Estimation

Table 6 and Table 7 show embarrassingly unsatisfactory results of question pattern identification. In the tables, P and GT in the Focus column indicate the cases where the predicted question foci and ground-truth are respectively used. As already discussed, these low per-
formances obviously affected the quality of generated questions.

| $N_c$ | Focus | $n$ | $N$ | Accuracy (%) |
|------|-------|-----|-----|--------------|
| 5    | P     | 1, 2, 3 | 200 | 0.45         |
| 10   | P     | 1, 2, 3 | 200 | 0.80         |
| 5    | GT    | 1, 2, 3 | 200 | 0.73         |
| 10   | GT    | 1, 2, 3 | 200 | 0.62         |

Table 6: Accuracy of question pattern classification.

| $N_c$ | Focus | $n$ | $B_1$ | $B_2$ | $B_3$ |
|------|-------|-----|------|------|------|
| 5    | P     | 20.00 | 3.39 | 0.000 |
| 10   | P     | 18.68 | 3.47 | 0.14  |
| 5    | GT    | 17.38 | 3.20 | 0.11  |
| 10   | GT    | 18.28 | 3.79 | 0.17  |

Table 7: Accuracy (BLEU scores) of question pattern generation.

Besides, the accuracies of generated question patterns are almost comparable across the predicted and the ground-truth question foci. This insists that the identification of question patterns is almost impossible by only relying on the current inputs (question focus and conversation history) and/or with the present models. This turns out that the process of question pattern identification has higher degree of freedom and should be more constrained with additional information such as entities appeared in the text passage.

### 6.4 Generated Question Examples

Figure 7 showcases generated examples.

In the top (good) example, both of question focus estimation and question pattern identification were correct, leading to the generation of a question that completely matched with the ground-truth question.

The second example exhibits a mixed case. As the generated question is largely different from the ground-truth question, the BLEU score is quite low. However, the generated question may be acceptable, given the QA conversation situation. This example suggests that we need to devise a better metrics for properly evaluating conversationally adequate questions.

The third and fourth examples present failed question generation cases. The former example shows failed question pattern identification and the latter example further exemplifies a fail in question pattern identification. As a result, the generated questions made no senses to the current question foci.

**Figure 7**: Good and bad examples of generated questions.

### 7 Conclusions

Conversational question generation (CQG) is a recently emerging area of NLP research initiated by (Gao et al., 2019). Given a range of potential practical applications, a question coherent to the current QA situation should be generated even without the corresponding answer provided. This study is first to propose a framework for answer-unaware CQG by assuming that the quality of questions can be improved by knowing the question focus and the question pattern. That is, the former contributes to choose a question topic (what-to-ask), and the later could lead the proper generation of the words in a question (how-to-ask). The experimental results confirmed that our research direction would be promising, but highlighted that further effort has to be made: in particular, the question pattern identification process should be greatly improved by enhancing the model and its ingredients.

To further push forward this new area of research, it would be necessary to establish a better evaluation metrics that could more adequately reflect the conversational natures of natural QA dialogues.
References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentai Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

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

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

Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. arXiv preprint arXiv:1905.08949.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.

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

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266.

Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. arXiv preprint arXiv:1706.02027.

Mark Yatskar. 2019. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

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

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. arXiv preprint arXiv:1704.01792.