Self-supervised Dialogue Learning for Spoken Conversational Question Answering

Nuo Chen\textsuperscript{1}, Chenyu You\textsuperscript{2}, Yuexian Zou\textsuperscript{1,3,*}

\textsuperscript{1}ADSPLAB, School of ECE, Peking University, Shenzhen, China
\textsuperscript{2}Department of Electrical Engineering, Yale University, CT, USA
\textsuperscript{3}Peng Cheng Laboratory, Shenzhen, China

\texttt{nuochen@pku.edu.cn, chenyu.you@yale.edu, zouyx@pku.edu.cn}

Abstract

In spoken conversational question answering (SCQA), the answer to the corresponding question is generated by retrieving and then analyzing a fixed spoken document, including multi-part conversations. Most SCQA systems have considered only retrieving information from ordered utterances. However, the sequential order of dialogue is important to build a robust spoken conversational question answering system, and the changes of utterances order may severely result in low-quality and incoherent corpora. To this end, we introduce a self-supervised learning approach, including incoherence discrimination, insertion detection, and question prediction, to explicitly capture the coreference resolution and dialogue coherence among spoken documents. Specifically, we design a joint learning framework where the auxiliary self-supervised tasks can enable the pre-trained SCQA systems towards more coherent and meaningful spoken dialogue learning. We also utilize the proposed self-supervised learning tasks to capture intra-sentence coherence. Experimental results demonstrate that our proposed method provides more coherent, meaningful, and appropriate responses, yielding superior performance gains compared to the original pre-trained language models. Our method achieves state-of-the-art results on the Spoken-CoQA dataset.

Index Terms: self-supervised learning, dialogue learning, spoken conversational question answering

1. Introduction

Spoken conversational question answering (SCQA) considers the problem of retrieving and aggregating spoken documents from a fixed collection, and then analyzes retrieved information to provide answers according to the given questions \cite{1}. In recent years, neural network based methods \cite{2, 3, 4, 5, 6} have attracted a lot of attention due to the advance in deep learning. Specifically, voice interfaces for dialogue systems have drawn massive attention of researchers. Several methods have been widely applied in various real-world scenarios, such as Microsoft XiaoIce and Apple Siri.

Prior SCQA systems focus on utilizing pre-trained language models (PLMs) as the backbone models and achieve promising results \cite{7, 8, 9}. However, building such systems still faces two major challenges: 1) most SCQA systems for spoken documents \cite{7, 8, 9} are basically composed of two systems (Automatic speech recognition (ASR) system and CQA system) to tackle these SCQA tasks. First, they utilize an ASR module to translate spoken contents into the corresponding transcripts. Then a Text-CQA module adopts them as input, and

\hspace{1cm}① Indicates equal contribution

\hspace{1cm}∗ Corresponding Author
In this study, we aim at engaging in conversation with humans to trackle this task. In the end, we present the proposed self-supervised dialogue learning tasks to explicitly capture coreference resolution, semantic relevance, and dialogue coherence. Specifically, the auxiliary tasks include three sub-tasks: incoherence discrimination, insertion detection, and question prediction. Then we jointly train a PLMs-based SCQA model in a multi-task manner to improve the model capability by learning task-specific knowledge to accurately predict answers while preserving dialogue coherence. Moreover, the proposed self-supervised tasks do not require additional annotation and thus can be widely applied in other downstream speech question answering tasks. Extensive experiments show that the proposed self-supervised dialogue learning tasks consistently bring remarkable performance improvements for state-of-the-art PLM-based SCQA models, and achieves superior results on Spoken-CoQA [1] dataset.

2. Methods

In this section, we first describe the task of spoken machine reading comprehension. Then we introduce how to apply PLMs to tackle this task. In the end, we present the proposed self-supervised dialogue learning method on modeling the inconsistent order detection, as shown in Figure 2.

2.1. Problem Formulation

In this study, we aim at engaging in conversation with humans in a dialogue-style interaction [14]. Given a SCQA dataset, we can formulate it as \( D = \{P_i, Q_i, A_i\} \), where \( P_i \) denotes the given passage, \( Q_i = \{q_i, q_i'_{-1}, ..., q_i'_{-L}\} \) and \( A_i = \{a_i, a_i'_{-1}, ..., a_i'_{-L}\} \) are the \( l \)-turn dialogue questions and the corresponding answers based on the \( \{P_i\} \), respectively. For example, given the passage and the corresponding dialogue question \( \{P_i, q_i\} \), the task is to provide the answer \( a_i \). The response depends on both the passage and history dialogue utterances. Inspired by recent success [15], we prepend the latest \( m \) rounds of dialogue to the current question \( q_i \) by incorporating conversation history for answer predictions. Consequently, the questions can be reformulated as \( q_i = \{q_i^{l-m}, q_i^{l-m-1}, ..., q_i^{l-1}, q_i^l\} \).

2.2. Pre-trained Language Models

Existing PLMs-based methods [16, 17, 18, 19] have achieved superior performance in a wide range of natural language processing tasks, such as question answering [20, 21, 22, 23] and text classification [24]. In this work, we build upon their success and evaluate our proposed method by incorporating it into BERT [16], which is the widely used PLMs-based architecture.

2.3. BERT for SCQA

We employ this PLMs-based approach in this work by applying BERT for SCQA tasks. The input sequence is obtained by concatenating auto-transcribed token sequence of dialogue questions and corresponding passage. In general, the input is formulated as,

\[
X = [\text{[CLS]}q_0^{l-1} \ldots q_i^l \text{[SEP]}P_i \text{[SEP]}],
\]

where “[CLS]” indicates the beginning token of every word sequence and “[SEP]” is a special separator token. Then the pretrained BERT is employed to extract hidden state features of each token. Next, a specific task layer is applied to predict an answer by incorporating conversation history for answer predictions. In detail, suppose that the \( k \)-th token and \( j \)-th token in the \( i \)-th
training examples are the start and end index of the answer span for \(l\)-th turn dialogue, SCQA answer prediction loss is given as:

\[
L_{ans} = - \sum_{l} \left[ \log(p_{i,l}^{k,\text{start}}) + \log(p_{i,l}^{k,\text{end}}) \right]
\]  

2.4. Self-Supervised Dialogue Learning Tasks

We aim at capturing dialogue coherent and consistency to produce more meaningful representations for performance improvements. Figure 2 illustrates the overview of the SCQA task and our proposed methods. Specifically, we design three auxiliary self-supervised tasks: incoherence discrimination, insertion detection, and question prediction. These tasks are jointly trained with the SCQA model in a multi-task manner. More concretely, we introduce two special tokens [INC], [INS] for two following downstream tasks, respectively.

2.4.1. Incoherence Discrimination

We argue that the sequential order of utterances within the dialogue is important for meaningful dialogue learning to solve the SCQA tasks. However, as noted in [17], BERT has the limitation in learning discourse-level semantic information due to NSP (next sentence prediction) [11]. In the context of SCQA, the model is required not only to discriminate whether the utterances are consecutive or not even though they are semantically related words, but also to distinguish the utterances with different semantic meanings.

To achieve this goal, we design incoherence discrimination. Specifically, we first extract \(k\) consecutive utterances from the dialogue context with respect to \(i\)-th passage, and then randomly delete one of them. In the training stage, we add [INC] tokens before each utterance to help the model find which is the deleted utterance. [INC] tokens refer the possible position of the deleted target utterance. Hence, the input for this sub-task can be given as:

\[
X_{\text{INC}} = [[\text{CLS}][\text{INC}]_{1}u_{1}^{i} [\text{INC}]_{2} \cdots u_{t-1}^{i} [\text{INC}]_{t}u_{t+1}^{i} \cdots [\text{INC}]_{k}u_{k}^{i} [\text{SEP}]u_{k}^{i} [\text{SEP}]],
\]

where \(u_{i}^{k}\) denotes vector concatenation of \(q_{i}^{k}, a_{i}^{k}\), and \(u_{i}^{t}\) is the deleted utterance.

2.4.2. Insertion Detection

Existing BERT-based models for multi-turn conversations have achieved superior performance in capturing discourse-level representations within the entire dialogue history. However, these models in multi-turn manner only learn token-level information, ignoring utterance-level interaction.

To enrich the utterance-level interaction between the utterances within a dialogue, we propose insertion detection. Formally, we first extract \(k\) consecutive utterances from \(i\)-th passage of the original dialogue contexts. Then we insert an utterance to a random dialogue from a set of \(k\) extracted utterances. For example, \(k + 1\) utterances consist of \(k\) utterances from original spoken documents and one from different corpus. The objective is to detect which one is the inserted among the \(k+1\) utterances. [INS] tokens are positioned before each utterance.

The input is given as follows:

\[
X_{\text{INS}} = [[\text{CLS}][\text{INS}]_{1}u_{1}^{i} [\text{INS}]_{2} \cdots [\text{INS}]_{t}u_{t}^{i,\text{ins}} [\text{INS}]_{t+1}u_{t+1}^{i} \cdots [\text{INS}]_{k+1}u_{k+1}^{i} [\text{SEP}]]
\]

where \(u_{i}^{t,\text{ins}}\) is the inserted utterance.

2.4.3. Question Prediction

In contrast to two previous auxiliary self-supervised tasks designed to capture dialogue consistency and coherence within a properly ordered dialog, we propose question prediction to learn temporal dependencies between semantically similar utterances within a dialogue. Given a dialogue history, \(C_{i} = \{u_{1}^{i}, u_{2}^{i}, \ldots, u_{t-1}^{i}\}\), randomly shuffled \(k\) questions, \(Q = \{q_{1}^{\text{end}}, q_{2}^{\text{end}}, \ldots, q_{k}^{\text{end}}\}\) and the passage \(P_{i}, u_{i}^{t}\) denotes the current corresponding question, the question prediction in this process aims to predict which question comes from a given context. Following the multi-choice question answering setting [13, 25], the input can be written as:

\[
X_{\text{QUE}} = [[\text{CLS}]C_{i}, Q_{m}[\text{SEP}]P_{i}[\text{SEP}]]
\]

where \(Q_{m}\) is the \(m\)-th question among \(Q\). In this process, to select a question candidate, we use a signal layer [26] to translate the hidden representation of ”[CLS]” for similarity scores. We then stack all the scores of \(k\) question candidates as input signal, and pass them to the softmax function. In this process, the selected tokens for each task (e.g., [INS], [INC], [INS]) are labeled as 1 or 0. We utilize binary cross-entropy loss as the optimization objective for these self-supervised tasks. Finally, the overall loss function is computed by combining SCQA answer prediction loss \(L_{ans}\) and auxiliary tasks loss with same ratio.

3. Experiments

3.1. Dataset

We evaluate our proposed method on Spoken-CoQA dataset. Spoken-CoQA [11] is an English spoken conversational question answering (SCQA) dataset, which consists of 40k question-answer pairs from 4k conversations in the training set and 380 conversations in the test set from five diverse domains, respectively. In Spoken-CoQA, questions are required to understand both passages and previous dialogue history, which is a challenging task to SCQA models. Note that questions and passages are both in text and spoken form, and answers are in text form, respectively. The word error rate (WER) is 18.7%.

3.2. Experimental Settings

We implement our model by using PyTorch. We adopt BERT-based as our backbone encoder, which consists of 12 transformer layers with maximum sequence length of the input set to 512 and the hidden vector dimension 768. The \(m\) and \(k\) are set to 2 and 9, respectively. In the training stage, the learning rate is set to \(3 \times 10^{-5}\) and batch size is 4 per GPU. We train each model on 2x 2080Ti for 2-3 days with 6 epochs. We utilize F1 and Exact Match (EM) for evaluating SCQA quality.

3.3. Results

We compare our proposed methods with five baselines\(^1\) in Table 1. We observe that our model significantly outperforms other baselines in most of the cases on Spoken-CoQA. This indicates that the proposed self-supervised learning tasks can explicitly enable the model to capture coherence and consistency within a dialogue. Compared with BERT, our method improves F1

---

\(^1\)For fair comparison, we do not compare our methods with baselines that have not been published.
score to 57.8% (vs. 54.1%). We also conduct ablation studies and report the results in Table 1. We see that sequentially add three proposed strategies bring significant performance improvements in F1 score. The results show that INC > QUE > INS, with respect to the importance of auxiliary manipulation strategy. This indicates that incorporating sequential information with the dialogue help to learn more contextual representations from the sequences. We also see that removing INS leads to the small performance drop (57.0% vs. 57.8%). This is because inserting a random question to some extent does not affect the discourse relations in the multi-turn conversations. Removing QUE, the F1 score for SCQA systems is dropped from 57.8% to 56.6%. This suggests that, by predicting which question comes from a given dialogue history and passage, the network can facilitate more interactions between passage and question for superior performance boosts.

4. Discussion

4.1. Dialogue Context History

In this study, we utilize dialogue context history via prepending the latest $m$ rounds of dialogue with the current question to incorporate contextual information, including questions and ground-truth answers (See Section 2.1). Here we investigate the effect of $m$, and report the results in Table 2. Specifically, we utilize BERT as the baseline. We observe that excluding dialogue history results in performance degradations by 5.6% and 7.8% in terms of EM and F1 score. This suggests that incorporating dialogue contextual information brings significant performance improvements. Meanwhile, we can see that when $m$ is set to 2, the model achieves the best performance.

4.2. Different Lengths of Dialogue Context

We conduct experiments to explore the impact of different lengths of dialogue turns on network performance. We report the experimental results of three auxiliary self-supervised tasks with different lengths of dialogue turns on Spoken-CoQA test set. The experimental results of BERT with three auxiliary tasks by examining all possible combinations in Figure 3. We can observe that the performance of training with three proposed self-supervised tasks shows a tendency to increase significantly ($\leq 12$). In addition, there exists the performance drop when the number of turns keeps increasing ($\geq 12$). This is because when trained with a few dialogues, the model cannot distinguish dialog consistency and coherence. On the other hand, when the number of average turns increase, rich dialogue history may introduce additional noisy signals, which significantly degrade the performance of the SCQA model. Based on the observations of training with three manipulation strategies, all of them show absolute improvements compared with the baseline. We achieve the best results when all the manipulation strategies are trained simultaneously (See red line in the Figure 3).

5. Conclusion

In this paper, we introduce a self-supervised dialogue learning for spoken conversational question answering by explicitly modeling dialogue coherence and consistency. Specifically, the model is fully trained in multi-task self-supervised manner and can be easily applied into existing state-of-the-art methods. Experimental results show that proposed methods achieve state-of-the-art performance on Spoken-CoQA dataset. In future, we expect our proposed self-supervised tasks in other language tasks, leading to further performance gains.

6. Acknowledgements

This paper is partially supported by Shenzhen municipal research funding project and Technology Fundamental Research Programs (No: GXWD20201231165807007-20200814115301001 & JSGG20191129105421211).

Table 1: Experimental results on the Spoken-CoQA test set. We use BERT-base as the baseline in this analysis. INC, INS, and QUE denote that the model trained with incoherence discrimination, insertion detection, and question prediction, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Child</th>
<th>Liter</th>
<th>Mid-High</th>
<th>News</th>
<th>Wiki</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>FlowQA [20]</td>
<td>35.8</td>
<td>35.2</td>
<td>34.2</td>
<td>33.6</td>
<td>34.7</td>
<td>34.7</td>
</tr>
<tr>
<td>BERT</td>
<td>55.6</td>
<td>54.6</td>
<td>52.7</td>
<td>53.8</td>
<td>53.8</td>
<td>54.1</td>
</tr>
<tr>
<td>BERT + SLU [27]</td>
<td>55.7</td>
<td>54.6</td>
<td>53.1</td>
<td>54.0</td>
<td>54.6</td>
<td>54.4</td>
</tr>
<tr>
<td>BERT + Sub-Word [12]</td>
<td>56.6</td>
<td>56.7</td>
<td>53.9</td>
<td>55.0</td>
<td>54.8</td>
<td>55.4</td>
</tr>
<tr>
<td>BERT + Cross-Attention [8]</td>
<td>57.1</td>
<td>56.1</td>
<td>56.0</td>
<td>54.6</td>
<td>55.0</td>
<td>55.6</td>
</tr>
<tr>
<td>BERT + INC</td>
<td>58.2</td>
<td>57.7</td>
<td>55.6</td>
<td>55.7</td>
<td>56.3</td>
<td>56.7</td>
</tr>
<tr>
<td>BERT + INS</td>
<td>57.2</td>
<td>56.7</td>
<td>55.6</td>
<td>55.2</td>
<td>55.3</td>
<td>56.0</td>
</tr>
<tr>
<td>BERT + QUE</td>
<td>58.0</td>
<td>57.4</td>
<td>55.6</td>
<td>55.4</td>
<td>55.6</td>
<td>56.4</td>
</tr>
<tr>
<td>BERT + INS + QUE</td>
<td>57.4</td>
<td>56.6</td>
<td>55.0</td>
<td>56.4</td>
<td>55.7</td>
<td>56.7</td>
</tr>
<tr>
<td>BERT + INC + QUE</td>
<td>58.1</td>
<td>57.8</td>
<td>56.3</td>
<td>56.8</td>
<td>56.2</td>
<td>57.2</td>
</tr>
<tr>
<td>BERT + INC + INS</td>
<td>58.2</td>
<td>57.6</td>
<td>55.4</td>
<td>56.0</td>
<td>56.1</td>
<td>56.9</td>
</tr>
<tr>
<td>BERT + INC + INS + QUE (Ours)</td>
<td>59.3</td>
<td>58.4</td>
<td>56.8</td>
<td>57.6</td>
<td>56.8</td>
<td>57.8</td>
</tr>
</tbody>
</table>

Table 2: Experimental results on Spoken-CoQA dataset when prepending different number of history questions and answers to the current question.

<table>
<thead>
<tr>
<th>Dialogue History rounds $m$</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>34.2</td>
<td>44.2</td>
</tr>
<tr>
<td>1</td>
<td>39.8</td>
<td>52.0</td>
</tr>
<tr>
<td>2</td>
<td>40.6</td>
<td>54.1</td>
</tr>
<tr>
<td>3</td>
<td>40.0</td>
<td>53.1</td>
</tr>
<tr>
<td>4</td>
<td>39.3</td>
<td>51.2</td>
</tr>
</tbody>
</table>

Figure 3: Experimental results of three auxiliary self-supervised tasks with different lengths of dialogue turns on Spoken-CoQA test set.
7. References


