INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Self-supervised Dialogue Learning for Spoken Conversational Question Answering

Nuo Chen[1†], *Chenyu You*[2†], *Yuexian Zou*[1,3,*]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Department of Electrical Engineering, Yale University, CT, USA
[3]Peng Cheng Laboratory, Shenzhen, China

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 [1]. In recent years, neural network based methods [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 [1, 7, 8]. However, building such systems still faces two major challenges: 1) almost any SCQA systems for spoken documents [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 provides the final prediction answer. An example from SCQA procedure is depicted in Figure 1. In general, ASR transcripts are not well-applicable to every spoken document. Thus, in this procedure, the network performance is mainly limited by the highly noisy data in ASR transcripts due to the word recognition errors (e.g., *poor* to *pool*), and lack of capitalisation and punctuation. 2) The utterance order in a meaningful and coherent dialogue largely determines the response generation accuracy [10, 11]. Existing SCQA systems rely heavily on contextual understanding in the dialogue flow. Thus, how to utilize the sequential order within spoken conversational documents as the self-supervisory signal to guide coherent dialogue learning is essential for the model developments.

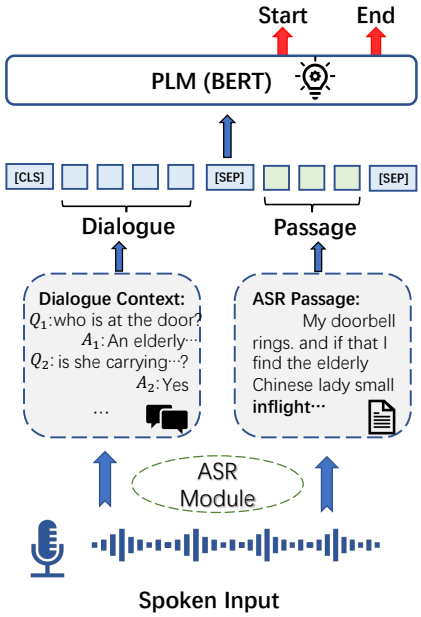

Figure 1: *The overview of the PLMs-based SCQA pipeline. "start" and "end" indicate the start and end indexes of prediction answers, respectively.*

To tackle the above-mentioned challenges, several studies have been proposed. Recent studies [9, 12] explored utilizing sub-word units to replace the whole word in order to reduce ASR errors. More recently, [13] combined speech content and text transcriptions based on PLMs to mine potentially useful clues in acoustic signals for better answer predictions, which may be influenced negatively by highly noisy ASR transcriptions. Most recently, [1, 8] introduced a teacher-student framework to distill incorruptible knowledge from a weighted teacher to make the student more ASR robust. However, none of existing works explicitly model the sequential order and evaluate its

† Indicates equal contribution
* Corresponding Author

http://dx.doi.org/10.21437/Interspeech.2021-120

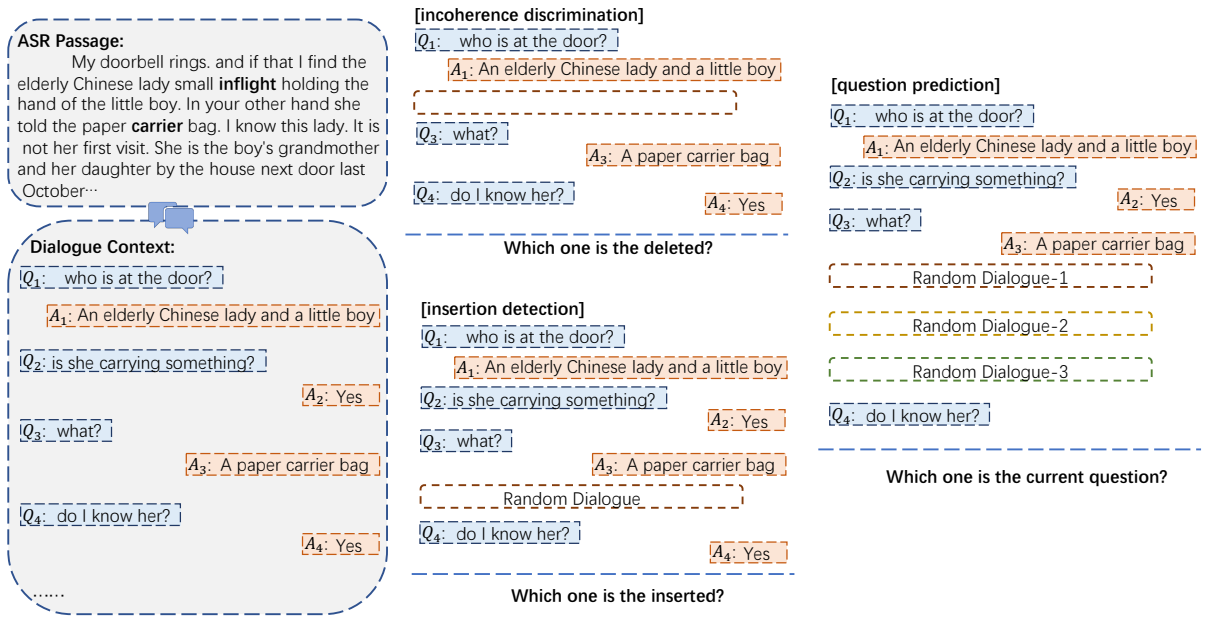

Figure 2: *An overview of our proposed self-supervised dialogue learning framework. The input sequence for each sub-task (k consecutive dialogue utterances) is dynamically selected from the original dialogue history within the fixed passage during training. For example, "Random Dialogue" is the utterance that is randomly extracted from an random dialogue.*

conversation criticality to the developments of SCQA systems.

In this paper, we work towards modeling dialogue flow within a fixed passage for the downstream SCQA tasks with performance improvements. The objective is achieved by exploring the coherence and consistency in a conversation as the useful self-supervised signals to guide spoken dialogue learning. In implementation, we introduce several 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. We then introduce how to apply PLMs to trackle 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 $\mathcal{D} = \{P_i, Q_i, A_i\}_1^N$, where P_i denotes the given passage. Q_i=$\{q_i^1, q_i^2, ..., q_i^L\}$ and A_i= $\{a_i^1, a_i^2, ..., a_i^l\}$ are l-turn dialogue questions and the corresponding answers based on the $\{P_i\}_1^N$, respectively. For example, given the passage and the corresponding dialogue question $\{P_i, q_i^l\}$, the task is to provide the answer a_i^l. 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^l by incorporating conversation history for answer predictions. Consequently, the questions can be reformulated as q_i^l= $\{q_i^{l-m}, a_i^{l-m}; ..; q_i^{l-1}, a_i^{l-1}\}$.

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 SCAQ 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,

$$\mathbf{X} = [[\mathrm{CLS}]\mathrm{q_i^1}\mathrm{a_i^1}...\mathrm{q_i^l}\mathrm{a_i^l}[\mathrm{SEP}]\mathrm{P_i}[\mathrm{SEP}]], \quad (1)$$

where "[CLS]" indicates the beginning token of every word sequence and "[SEP]" is a special separator token. Then the pre-trained BERT is employed to extract hidden state features of each token. Next, a specific task layer is applied to predict an emphasis probability distribution over these representations. Finally, we adopt the cross-entropy loss as the training objective. 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:

$$\mathcal{L}_{ans} = -\sum_{i}\sum_{l}[log(p_{i,l}^{k,start}) + log(p_{i,l}^{j,end})] \quad (2)$$

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:

$$\begin{aligned}\mathbf{X_{INC}} = [[\mathrm{CLS}][\mathrm{INC}]_1 \mathrm{u}_\mathrm{i}^1 [\mathrm{INC}]_2 \cdots \mathrm{u}_\mathrm{i}^{\mathrm{t}-1} [\mathrm{INC}]_\mathrm{t} \\ \mathrm{u}_\mathrm{i}^{\mathrm{t}+1} \cdots [\mathrm{INC}]_\mathrm{k} \mathrm{u}_\mathrm{i}^\mathrm{k} [\mathrm{SEP}] \mathrm{u}_\mathrm{i}^\mathrm{t} [\mathrm{SEP}]],\end{aligned} \quad (3)$$

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 dialog history. However, these model in multi-turn manner only learn token-level information [17], 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:

$$\begin{aligned}\mathbf{X_{INS}} = [[\mathrm{CLS}][\mathrm{INS}]_1 \mathrm{u}_\mathrm{i}^1 [\mathrm{INS}]_2 \cdots [\mathrm{INS}]_\mathrm{t} \mathrm{u}_\mathrm{i}^{\mathrm{ins}} \\ [\mathrm{INS}]_{\mathrm{t}+1} \mathrm{u}_\mathrm{i}^\mathrm{t} \cdots [\mathrm{INS}]_{\mathrm{k}+1} \mathrm{u}_\mathrm{i}^\mathrm{k} [\mathrm{SEP}]]\end{aligned} \quad (4)$$

where u_i^{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, $C_i = \{u_i^1, u_i^2, .., u_i^{t-1}\}$, randomly shuffled k questions, $Q = \{q_1^{rand}, q_2^{rand}, u_i^t, .., q_k^{rand}\}$ 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:

$$\mathbf{X_{QUE}} = [[\mathrm{CLS}]\mathrm{C_i}, \mathrm{Q_m}[\mathrm{SEP}]\mathrm{P_i}[\mathrm{SEP}] \quad (5)$$

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., $[\mathrm{INS}]_\mathrm{t}, [\mathrm{INC}]_\mathrm{t}$) 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 $\mathcal{L}_{ans}$ and auxiliary tasks loss with same ratio.

3. Experiments

3.1. Dataset

We evaluate our proposed method on Spoken-CoQA dataset. Spoken-CoQA [1] 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 3×10^{-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.

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.*

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

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

Dialogue History rounds m	EM	F1
0	34.2	44.2
1	39.8	52.0
2	**40.6**	**54.1**
3	40.0	53.1
4	39.3	51.2

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

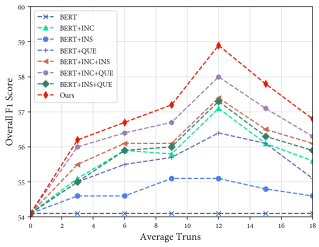

Figure 3: *Experimental results of three auxiliary self-supervised tasks with different lengths of dialogue turns on Spoken-CoQA test set.*

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).

7. References

[1] C. You, N. Chen, F. Liu, D. Yang, and Y. Zou, "Towards data distillation for end-to-end spoken conversational question answering," *arXiv preprint arXiv:2010.08923*, 2020.

[2] C. You, R. Zhao, L. Staib, and J. S. Duncan, "Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation," *arXiv preprint arXiv:2105.07059*, 2021.

[3] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong *et al.*, "CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (gan-circle)," *IEEE Transactions on Medical Imaging*, 2019.

[4] C. You, J. Yang, J. Chapiro, and J. S. Duncan, "Unsupervised wasserstein distance guided domain adaptation for 3d multi-domain liver segmentation," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*.

[5] C. You, Q. Yang, H. Shan, L. Gjesteby, G. Li, S. Ju, Z. Zhang, Z. Zhao, Y. Zhang, W. Cong *et al.*, "Structurally-sensitive multi-scale deep neural network for low-dose ct denoising," *IEEE Access*, 2018.

[6] C. You, L. Yang, Y. Zhang, and G. Wang, "Low-dose ct via deep cnn with skip connection and network-in-network," in *Developments in X-Ray Tomography XII*, 2019.

[7] C. You, N. Chen, and Y. Zou, "Knowledge distillation for improved accuracy in spoken question answering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[8] C. You, N. Chen, and Y. Zou, "Contextualized attention-based knowledge transfer for spoken conversational question answering," *arXiv preprint arXiv:2010.11066*, 2020.

[9] C.-H. Lee, S.-M. Wang, H.-C. Chang, and H.-Y. Lee, "ODSQA: Open-domain spoken question answering dataset," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 949–956.

[10] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan, "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*, 2016, pp. 110–119.

[11] J. Wu, X. Wang, and W. Y. Wang, "Self-supervised dialogue learning," *arXiv preprint arXiv:1907.00448*, 2019.

[12] C.-H. Li, S.-L. Wu, C.-L. Liu, and H.-y. Lee, "Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension," *arXiv preprint arXiv:1804.00320*, 2018.

[13] C. Kuo, S. Luo, and K. Chen, "An audio-enriched bert-based framework for spoken multiple-choice question answering," in *INTERSPEECH*. ISCA, 2020, pp. 4173–4177.

[14] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1371–1374.

[15] C. Zhu, M. Zeng, and X. Huang, "SDNet: Contextualized attention-based deep network for conversational question answering," *arXiv preprint arXiv:1812.03593*, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," in *International Conference on Learning Representations*, 2020.

[18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5753–5763.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[20] H.-Y. Huang, E. Choi, and W.-t. Yih, "FlowQA: Grasping flow in history for conversational machine comprehension," *arXiv preprint arXiv:1810.06683*, 2018.

[21] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[22] H.-Y. Huang, C. Zhu, Y. Shen, and W. Chen, "Fusionnet: Fusing via fully-aware attention with application to machine comprehension," *arXiv preprint arXiv:1711.07341*, 2017.

[23] N. Chen, F. Liu, C. You, P. Zhou, and Y. Zou, "Adaptive bi-directional attention: Exploring multi-granularity representations for machine reading comprehension," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[24] J. Chen, Z. Yang, and D. Yang, "Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification," in *ACL*. Association for Computational Linguistics, 2020, pp. 2147–2157.

[25] M. Yan, H. Zhang, D. Jin, and J. T. Zhou, "Multi-source meta transfer for low resource multiple-choice question answering," in *ACL*. Association for Computational Linguistics, 2020, pp. 7331–7341.

[26] T. Whang, D. Lee, D. Oh, C. Lee, K. Han, D. Lee, and S. Lee, "Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection," *CoRR*, vol. abs/2009.04703, 2020.

[27] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5754–5758.