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

Contextualized Attention-based Knowledge Transfer for Spoken Conversational Question Answering

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

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

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

Abstract

Spoken conversational question answering (SCQA) requires machines to model the flow of multi-turn conversation given the speech utterances and text corpora. Different from traditional text question answering (QA) tasks, SCQA involves audio signal processing, passage comprehension, and contextual understanding. However, ASR systems introduce unexpected noisy signals to the transcriptions, which result in performance degradation on SCQA. To overcome the problem, we propose CADNet, a novel contextualized attention-based distillation approach, which applies both cross-attention and self-attention to obtain ASR-robust contextualized embedding representations of the passage and dialogue history for performance improvements. We also introduce the spoken conventional knowledge distillation framework to distill the ASR-robust knowledge from the estimated probabilities of the *teacher* model to the *student*. We conduct extensive experiments on the Spoken-CoQA dataset and demonstrate that our approach achieves remarkable performance in this task.

Index Terms: spoken conversational question answering, machine reading comprehension, conversational question answering

1. Introduction

Neural network based end-to-end methods [1, 2, 3, 4, 5] have attracted a lot of attention in the machine learning community. With the recent advances in machine learning, spoken question answering (SQA) has become an important research topic during the past few years. To be specific, SQA requires the machine to fully understand the spoken content of the document and questions, and then predict an answer. A major limitation for this field is the lack of benchmark datasets. To alleviate such issue, several benchmark datasets [6, 7, 8, 9] are released to the speech processing and natural language processing communities. Spoken-SQuAD [7] is one of the typical benchmarks, which uses CMU Sphinx to generate auto-transcribed text given the Text-SQuAD [10] dataset. However, the SQA tasks only share the single-turn setting to answer a single question given the spoken document, which is far from real SQA scenarios. For example, in a real-world interview, the discourse structure is more complex, which including multi-part conversation. Thus, it is crucial to enable the QA systems to address the turn structure (e.g., meetings, debate) and the consistent discourse interpretation.

SQA includes audio signal processing, passage comprehension, and contextual understanding. The typical way is to translate speech content into text forms, and then apply the state-of-the-art Text-QA models on the transcribed text documents [8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 19, 22]. However, the recent study [23] suggests that ASR errors may severely affect answering accuracy. Previous works have been proposed to address such issues. Lee et al. [8] shows that using sub-word units yields promising improvements in terms of accuracy. Very recently, domain adversarial learning [12] was introduced to mitigate the effects of ASR errors effectively. Most recently, You et al. [9] released the first SCQA benchmark dataset - Spoken-CoQA, and then propose to use a *teacher-student* paradigm to boost the network performance on highly noisy ASR transcripts.

In this paper, based on our previous work [9] on the spoken conversational question answering task, we present CADNet, a contextual attention-based data distillation network. Our network first leverages both cross-attention and self-attention to extract relevant information between auto-transcribed (ASR) texts and the reference text documents to better understand the corpus of spoken documents and questions effectively. Then

Table 1: *An example from Spoken-CoQA. We can observe large misalignment between the manual transcripts and the corresponding ASR transcripts. Note that the misalignment is in* ***bold*** *font.*

Manual Transcript	**ASR Transcript**
Once there was a beautiful fish named Asta. Asta lived in the ocean. There were lots of other fish in the ocean where Asta lived. They played all day long. One day, a bottle floated by over the heads of Asta and his friends. They looked up and saw the bottle ...	once there was a beautiful fish named **After**. **After** lived in the ocean. There were lots of other fish in the ocean **we're** asked to live. They played all day long. One day, a bottle floated by over the heads of **vast** and his friends. They looked up and saw the bottle ...
Q_1: What was the name of the fish? A_1: Asta R_1:Asta.	ASR-Q_1: What was the name of the fish? A_1: **After** R_1: **After**.
Q_2: What looked like a birds belly? A_2: a bottle R_2: a bottle	ASR-Q_2: What looked like a bird to **delhi**? A_2: a bottle R_2: a bottle

† Indicates equal contribution
* Corresponding Author

http://dx.doi.org/10.21437/Interspeech.2021-110

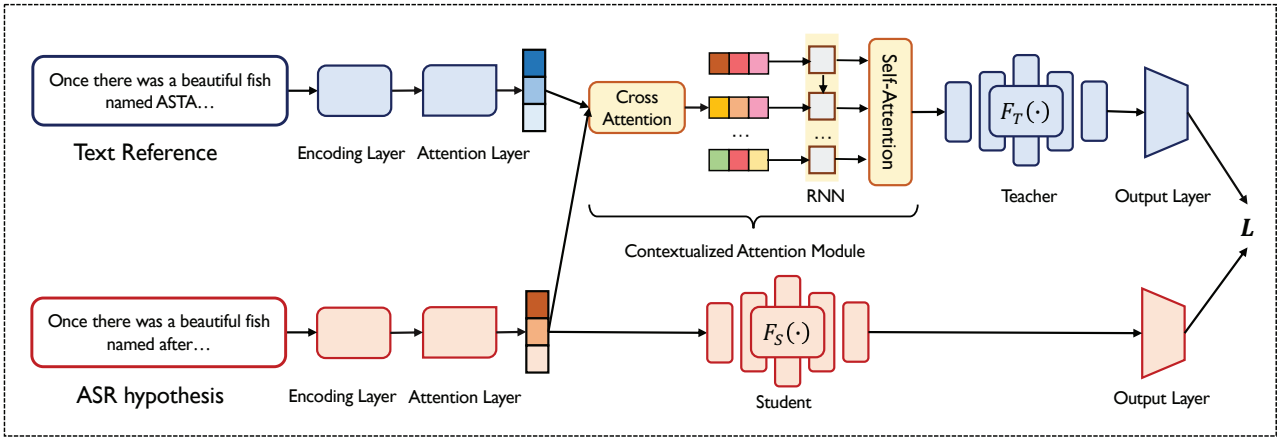

Figure 1: *Overview of our proposed method.*

we distill reliable supervision signals from the reference written documents, and uses these predictions to guide the training of the *student*. We evaluate our method on the Spoken-CoQA dataset, and experimental results demonstrate that our method exhibits good improvements in terms of accuracy over several state-of-the-art models.

2. Dataset

We use the listening comprehension benchmark dataset Spoken-CoQA [9] corpus. Each example in this dataset is defined as follows: $\{P_i, Q_i, A_i\}_1^N$, where P_i denotes the given passage. Q_i=$\{q_{i1}, q_{i2}, ..., q_{iL}\}$ and A_i= $\{a_{i1}, a_{i2}, ..., a_{iL}\}$ represent a passage with L-turn queries and corresponding answers, respectively. Given a passage P_i and multi-turn history questions $\{q_{i1}, q_{i2}, ..., q_{iL-1}\}$ and answers $\{a_{i1}, a_{i2}, ..., a_{iL-1}\}$, our goal is to generate a_{iL} for the given current question q_{iL}. Note that questions and documents in Spoken-CoQA are in both text and spoken forms, and answers are in the text form. Table 1 is an example selected from the Spoken-CoQA development set. As we can see, given the text document (ASR-document) the conversation begins with the question Q_1 (ASR-Q_1), then the Text-QA model needs to answer Q_1 (ASR-Q_1) with A_1 based on a contiguous text span R_1. This suggests that ASR transcripts (both the document and questions) are more difficult for the model to comprehend, reason and even predict correct answers. The word error rate (WER) is 18.7%.

3. Method

3.1. Model Overview

In this work, we focus primarily on applying the existing Text-based Conversational Machine Reading Comprehension (Text-CMRC) models to handle highly noisy transcriptions from the ASR systems. Generally, the Text-CMRC models include three major parts: encoding layer, attention layer, and output layer. *Encoding Layers* uses the documents and conversations (questions and answers) to encode the word into the corresponding feature embedding (e.g., character embedding, word embedding, and contextual embedding). In the *Attention Layer*, we then extract the most relevant information from the context for answering the question by condensing the context representations of documents into a fixed-length vector. Finally, *Output Layer* predicts an answer for the current question given the learned representations.

To mitigate the adverse effects caused by the ASR errors in the SCQA task, we propose a novel contextualized attention-based distillation network (CADNet). We first introduce the cross-attention mechanism to align the mismatch between the reference manual transcript and the corresponding ASR transcripts. We then use multi-layer recurrent neural networks (RNN) to learn ASR-robust contextual representations from the document collections. Next, we conduct self-attention to establish the correlations between words at different positions and capture additional cues from the manual transcripts. Finally, we improve the knowledge transfer by training a robust *teacher* model and then enroll the improved knowledge into *student* model trained on ASR transcriptions to provide better performance improvements. We present the proposed method in Figure1.

3.2. Contextualized Attention (CA) Module

Cross-Attention We conduct cross-attention to capture the relevance between the reference written documents and the corresponding ASR transcriptions. Given the representations of the reference text documents and the corresponding ASR transcriptions with n tokens: $\{\mathbf{w}_1^T, ..., \mathbf{w}_n^T\} \subset \mathbf{R}^d$ and $\{\mathbf{w}_1^A, ..., \mathbf{w}_n^A\} \subset \mathbf{R}^d$, the attention function is computed as:

$$H_{ij} = ReLU(S\mathbf{w}_i^T)DReLU(S\mathbf{w}_j^A) \quad (1)$$

$$\theta_{ij} \propto Exp(H_{ij}) \quad (2)$$

$$\hat{\mathbf{w}}_i^T = \sum_j \theta_{ij}\mathbf{w}_j^A \quad (3)$$

where $S \in \mathbf{R}^{d\times k}$, $D \in \mathbf{R}^{k\times k}$ is a diagonal matrix, and k denotes the attention hidden size. For brevity, we re-formulate the above word-level attention function as $\mathbf{Attn}(\{\mathbf{w}_i^T\}_{i=1}^n, \{\mathbf{w}_i^A\}_{i=1}^n, \{\mathbf{w}_i^A\}_{i=1}^n)$ (See in Figure 2).

RNN In order to obtain the better ASR-robust contextualized representations, we use the BiLSTM after the cross attention layer:

$$\tilde{\mathbf{w}}_1^{T,L+1}, ..., \tilde{\mathbf{w}}_n^{T,L+1} = BiLSTM(\hat{\mathbf{w}}_1^T, ..., \hat{\mathbf{w}}_n^T) \quad (4)$$

$$\hat{\mathbf{w}}_i^T = [\hat{\mathbf{w}}_1^{T,1}; ...; \hat{\mathbf{w}}_1^{T,L}] \quad (5)$$

Table 2: *Comparison of four baselines. Note that we denote Text-CoQA and Spoken-CoQA test set as T-CoQA and S-CoQA test for brevity.*

	T-CoQA				S-CoQA			
	T-CoQA dev		S-CoQA test		T-CoQA dev		S-CoQA test	
Methods	EM	F1	EM	F1	EM	F1	EM	F1
FlowQA [24]	66.8	75.1	44.1	56.8	40.9	51.6	22.1	34.7
SDNet [25]	68.1	76.9	39.5	51.2	40.1	52.5	41.5	53.1
BERT-base [26]	67.7	77.7	41.8	54.7	42.3	55.8	40.6	54.1
ALBERT-base [27]	71.4	80.6	42.6	54.8	42.7	56.0	41.4	55.2
Average	68.5	77.6	42	54.4	41.5	54.0	36.4	49.3

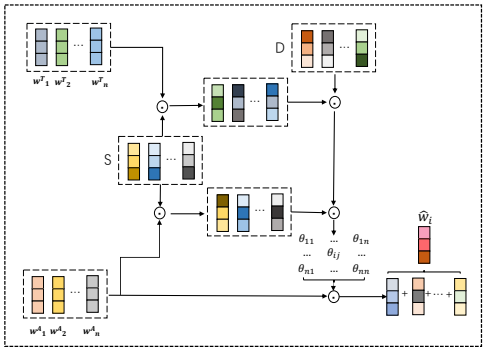

Figure 2: *Architecture for Cross Attention Mechanism.* $\odot$ *is the element-wise multiplication.*

where L denotes the number of RNN layers.

Self-Attention Considering there exists a huge mismatch between ASR transcripts and the corresponding text documents, it is necessary to establish direct correlations between two transcriptions. In order to model the long-term dependency between two documents, we additionally introduce a self-attention layer to obtain self-attentive representations. The self-attentive vector can be computed as:

$$\{\mathbf{u}_i\}_{i=1}^n = \mathbf{Attn}(\{\tilde{\mathbf{w}}_i^T\}_{i=1}^n, \{\tilde{\mathbf{w}}_i^T\}_{i=1}^n, \{\tilde{\mathbf{w}}_i^T\}_{i=1}^n). \quad (6)$$

3.3. Knowledge Distillation

Inspired by the previous work [28] that trains a *student* model to match the full softmax distribution of the *teacher* model, we adopt a *teacher-student* paradigm to distill ASR-robust knowledge into a single *student* CMRC model.

Let $z_T = \{\mathbf{u}_i\}_{i=1}^n$ and $z_S = \{\mathbf{w}_i^A\}_{i=1}^n$ denote the training input to the following *teacher* model and *student* model, respectively. $y = \{\mathbf{A}_i\}_{i=1}^n$ denote corresponding sequence of gold labels. $\Psi_S = F_S(z_S)$ and $\Psi_T = F_T(z_T)$ are output score of potential *teacher* and *student* function, respectively. The final loss of the *student* is defined as:

$$L = \alpha L_{NLL}(p_\tau(\Psi_S), p_\tau(\Psi_T)) + (1-\alpha) L_{KD}(\Psi_T, y)), \quad (7)$$

where L_{NLL} and L_{KD} are the negative log-likelihood loss and cross entropy loss, respectively. $p_\tau(\cdot)$ denotes the softmax function [29]. τ and α are hyperparameters.

4. Experiments

In this section, we first introduce several previous state-of-the-art language models. Then we investigate the performance of all evaluated baselines trained on Text-CoQA [30] or Spoken-CoQA dataset. Finally, we demonstrate the effectiveness of the proposed CADNet.

4.1. Baseline Models

FlowQA [24] utilizes high-level granularity representations of questions and documents to enable the model to comprehend the topic of documents and integrate the underlying semantics of the dialogue history.

SDNet [25] uses inter-attention and self-attention to extract different levels of granularities to enable the model to understand the relevant information from passages and innovatively incorporate the BERT model.

BERT-base [26] is a remarkable breakthrough in the natural language processing community, which achieves state-of-the-art performances in many downstream natural language processing tasks. It utilizes stacked transformers as encoders with residual structures.

ALBERT-base [27], a lightweight variant of BERT-based models, uses the parameter-sharing strategy in multiple parts of a model to compress the model size. It shares a similar structure with BERT, while maintaining comparable performances with smaller parameters.

4.2. Experimental Settings

We choose BERT-base and ALBERT-base in this study, consisting of 12 transformer encoders with hidden size 768. To guarantee the integrity of training, we follow the standard settings in four baselines. BERT and ALBERT utilize BPE as the tokenizer, but FlowQA and SDNet use SpaCy [32] for tokenization. Specifically, in the case that each token in spaCy corresponds to more than one BPE sub-tokens, we compute the embedding for each token by averaging the BERT embeddings of the relevant BPE sub-tokens. Note that we train all the evaluated baseline models over Text-CoQA in our local computing environment, which results in a little bit different from their results on the Text-CoQA leaderboard. In detail, we train FLowQA and SDNet for 30 epochs and fine-tune BERT and ALBERT for 5 epochs. We empirically set α to 0.9 (defined in Section 3.3) in all experiments. We choose the maximum EM (Exact Match) and F1 score for evaluating the performance of SCQA models.

4.3. Results

In this study, we choose four state-of-the-art CMRC models (FlowQA [24], SDNet [25], BERT-base [26], ALBERT-base [27]). We conduct two sets of experiments: 1). We train the baselines on Text-CoQA training set, and then evaluate the

Table 3: *Comparison of model performance. We set the model on text corpus as the teacher model, and the another on the ASR transcripts as the student model.*

	T-CoQA dev		S-CoQA test	
Methods	EM	F1	EM	F1
FlowQA [24]	40.9	51.6	22.1	34.7
FlowQA [24]+ sub-word unit [7]	41.9	53.2	23.3	36.4
FlowQA [24]+ SLU [31]	41.2	52.0	22.4	35.0
FlowQA [24]+**CA**	41.6	52.6	23.2	35.8
FlowQA [24]+**KD**	42.5	53.7	23.9	39.2
FlowQA [24]+**CA+KD**	**43.2**	**54.7**	**25.0**	**40.3**
SDNet [25]	40.1	52.5	41.5	53.1
SDNet [25]+ sub-word unit [7]	41.2	53.7	41.9	54.7
SDNet [25]+ SLU [31]	40.2	52.9	41.7	53.2
SDNet [25]+**CA**	40.9	54.1	42.4	54.1
SDNet [25]+**KD**	41.7	55.6	43.6	56.7
SDNet [25]+**CA+KD**	**42.5**	**57.2**	**44.5**	**57.7**
BERT-base [26]	42.3	55.8	40.6	54.1
BERT-base [26]+ sub-word unit [7]	43.2	56.8	41.6	55.4
BERT-base+ SLU [31]	42.5	56.1	41.0	54.6
BERT-base [26]+**CA**	43.3	56.6	41.6	55.2
BERT-base [26]+**KD**	44.1	58.8	42.8	57.7
BERT-base [26]+**CA+KD**	**45.1**	**59.9**	**43.8**	**58.9**
ALBERT-base [27]	42.7	56.0	41.4	55.2
ALBERT-base [27]+ sub-word unit [7]	43.7	57.2	42.6	56.8
ALBERT-base [27]+ SLU [31]	42.8	56.3	41.7	55.7
ALBERT-base [27]+**CA**	43.8	57.7	42.2	56.2
ALBERT-base [27]+**KD**	44.8	59.6	43.9	58.7
ALBERT-base [27]+**CA+KD**	**45.9**	**61.3**	**44.7**	**59.7**

baselines on Text-CoQA dev set and Spoken-CoQA dev set, respectively; 2). We train the baselines on Spoken-CoQA training set and compare the baselines on Text-CoQA dev set and Spoken-CoQA test set, respectively.

In the first set of experiments, we report the results in Table 2. When looked into the table, we can find a large performance gap between the model trained on the text references (Text-CoQA training set) and another on ASR transcriptions (Spoken-CoQA training set), which indicates that recognition errors inevitably misled the CMRC models to make incorrect predictions. Therefore, it is desirable and meaningful to explore a more appropriate strategy to alleviate the adverse effects of ASR errors.

In the second set of experiments, Table 3 reports the quantitative results. Our proposed *teacher-student* paradigm generally helps improve the baseline performance by injecting ASR-robust features into the baseline. With CA+KD, FlowQA achieves 54.7% (vs.51.6%), and 39.9% (vs.34.7%) on F1 score over the manual transcripts and ASR transcripts, respectively; SDNet outperforms the baseline model, achieving 56.5% (vs.52.5%) and 57.7% (vs.53.1%) on F1 score. As for two BERT-like models: BERT-base and ALBERT-base, our proposed KD strategy consistently leads to improvements over directly using baseline models, which achieves the F1 score of 58.8% (vs.55.8%) and 57.7% (vs.54.1%); 59.6% (vs.56.0%) and 58.7% (vs.55.2%). This demonstrates the effectiveness of our proposed KD strategy. We also evaluate the robustness of our CA mechanism. As shown in Table 3, our models trained with our CA mechanism further improve the results considerably. This suggests that utilizing CA can help the model comprehend the conversational context and extract relevant information from the passage, which is beneficial for final answer prediction.

To further illustrate the generalization of the proposed approach, we conduct experiments to investigate model performance on speech documents with different Word Error Rates

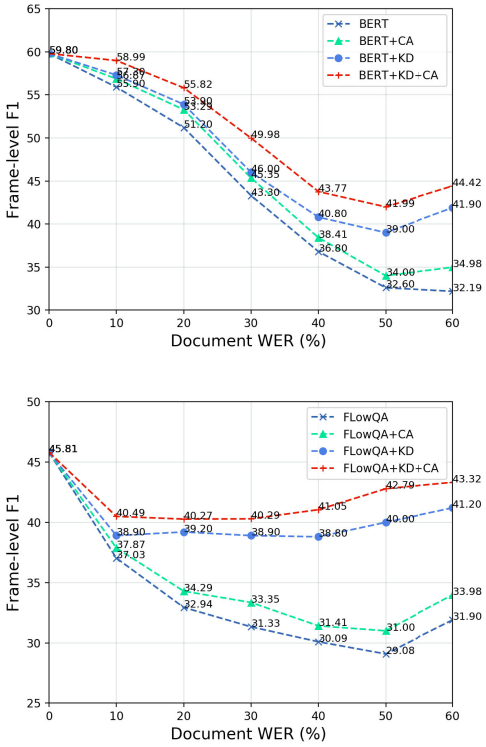

Figure 3: *Analysis of different WER on Spoken-CoQA.*

(WERs). We first split Spoken-CoQA into small sub-sets with different WERs. We then utilize Frame-level F1 score [33] as an additional evaluation metric to validate our proposed method on Spoken-CoQA. In Figure 3, we can learn that all the methods (FlowQA and BERT) have a significant drop in performance at higher WER. We observe that sequentially using the knowledge distillation strategy and contextualized attention mechanism are capable of yielding considerable performance boosts on all the evaluated models. This suggests that adopting these two strategies on SCQA tasks is able to improve network performance at higher WER.

5. Conclusion

In this paper, we propose a contextual attention-based knowledge distillation network, CADNet, to tackle the spoken conversational question answering task. We propose to leverage cross-attention and self-attention on both manual and ASR transcriptions, and employ a *teacher-student* framework to distill ASR-robust contextualized knowledge into the *student* model. Experimental results show that our method outperforms all the evaluated baselines and further mitigates the negative effects of ASR errors. Our future work is to explore more effective speech and text fusing mechanisms to further improve performance.

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, 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.

[2] 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.

[3] 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*.

[4] 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.

[5] 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.

[6] L.-s. Lee, J. Glass, H.-y. Lee, and C.-a. Chan, "Spoken content retrieval—beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1389–1420, 2015.

[7] 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.

[8] 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.

[9] 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.

[10] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[11] S.-B. Luo, H.-S. Lee, K.-Y. Chen, and H.-M. Wang, "Spoken multiple-choice question answering using multimodal convolutional neural networks," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 772–778.

[12] C.-H. Lee, Y.-N. Chen, and H.-Y. Lee, "Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7300–7304.

[13] D. Su and P. Fung, "Improving spoken question answering using contextualized word representation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8004–8008.

[14] 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.

[15] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Florence, Italy: Association for Computational Linguistics, 7 2019.

[16] 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.

[17] Z. Sun, P. K. Sarma, W. A. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 8992–8999.

[18] M. Ünlü and E. Arisoy, "Uncertainty-aware representations for spoken question answering," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[19] D. Peskov, J. Barrow, P. Rodriguez, G. Neubig, and J. Boyd-Graber, "Mitigating noisy inputs for question answering," *arXiv preprint arXiv:1908.02914*, 2019.

[20] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," *arXiv preprint arXiv:2008.06682*, 2020.

[21] N. Chen, C. You, and Y. Zou, "Self-supervised dialogue learning for spoken conversational question answering," *arXiv preprint arXiv:2106.02182*, 2021.

[22] E. Arısoy and M. Ünlü, "Uncertainty-aware representations for spoken question answering," 2021.

[23] B.-H. Tseng, S.-S. Shen, H.-Y. Lee, and L.-S. Lee, "Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine," *arXiv preprint arXiv:1608.06378*, 2016.

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

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

[26] 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.

[27] 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.

[28] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[29] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, 2017, pp. 1321–1330.

[30] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 249–266, 2019.

[31] 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.

[32] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[33] Y.-S. Chuang, C.-L. Liu, and H.-Y. Lee, "SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering," *arXiv preprint arXiv:1910.11559*, 2019.