Dealing with Data Scarcity in Spoken Question Answering

Merve Ünlü Menevşe[1], Yusufcan Manav[1], Ebru Arisoy[2], Arzucan Özgür[1]
[1] Boğaziçi University, [2]MEF University
Istanbul, Turkey
{merve.unlu1, yusufcan.manav}@boun.edu.tr,
saraclare@mef.edu.tr, arzucan.ozgur@boun.edu.tr

Abstract

This paper focuses on dealing with data scarcity in spoken question answering (QA) using automatic question-answer generation and a carefully selected fine-tuning strategy that leverages limited annotated data (paragraphs and question-answer pairs). Spoken QA is a challenging task due to using spoken documents, i.e., erroneous automatic speech recognition (ASR) transcriptions, and the scarcity of spoken QA data. We propose a framework for utilizing limited annotated data effectively to improve spoken QA performance. To deal with data scarcity, we train a question-answer generation model with annotated data and then produce large amounts of question-answer pairs from unannotated data (paragraphs). Our experiments demonstrate that incorporating limited annotated data and the automatically generated data through a carefully selected fine-tuning strategy leads to 5.5% relative F1 gain over the model trained only with annotated data. Moreover, the proposed framework is also effective in high ASR errors.

Keywords: spoken question answering, question generation

1. Introduction

Spoken QA is the task of finding relevant answers to questions from spoken documents. In spoken QA, cascade and end-to-end models have been investigated. Cascade models first utilize ASR to obtain transcriptions of spoken documents, and then perform QA on those documents (Tseng et al., 2016; Ünlü et al., 2019; Lee et al., 2019b; Ünlü and Arisoy, 2021; Li et al., 2018; You et al., 2021). End-to-end models leverage acoustic and text data together to optimize QA performance (Chuang et al., 2019; Lin et al., 2022). Despite remarkable progress in text-based QA, spoken QA models face a distinct set of challenges that hinder their performance. These challenges arise from using spoken language, i.e., erroneous ASR transcriptions, as well as lack of large amounts of spoken annotated data resulting in data scarcity in spoken QA. As a result, spoken QA models often struggle to achieve comparable levels of accuracy and effectiveness observed in text-based QA models.

In this paper, we focus on dealing with data scarcity in spoken QA. We utilize automatic question generation (QG) to increase QA training data. QG automatically generates question-answer pairs from unannotated data that contain only short passages. Since QG may also generate inaccurate pairs, we propose a fine-tuning strategy to effectively incorporate the noisy data coming from QG into QA training. The proposed fine-tuning strategy leverages the limited annotated data and the automatically generated data, and improves the spoken QA performance in various limited annotated data settings and noisy acoustic conditions.

The contributions of this paper are as follows: (i) To the best of our knowledge, the effect of limited annotated data on spoken QA performance with a comprehensive set of experiments is presented for the first time. (ii) Automatic question generation is used to deal with the data scarcity problem in spoken QA. Even though question generation has been investigated for text-based QA, to the best of our knowledge, it has not been explored for spoken QA before. (iii) A fine-tuning strategy is proposed to effectively incorporate the automatically generated question-answer pairs into QA training. Our experiments yield significant improvements in limited data spoken QA, showing the effectiveness of our proposed approach, and reveal the importance of dealing with data scarcity in spoken QA.

2. Related Work

Spoken QA is a more challenging task than text-based QA due to using acoustic data, i.e., noisy ASR transcriptions in cascade spoken QA systems, and due to lack of spoken QA datasets. Several studies have shown the negative impact of ASR errors on spoken QA models (Ünlü et al., 2019; Li et al., 2018; Ünlü and Arisoy, 2021; Lee et al., 2019a). Various techniques have been developed to address this challenge in cascade models, including domain adaptation (Lee et al., 2019a), integrating confidence scores (Ünlü and Arisoy, 2021), and knowledge distillation (You et al., 2021). However, these methods are limited by the availability of annotated data, which is often scarce for spoken QA. Collecting data for spoken QA is time-consuming and expensive, as it requires large amounts of audio data annotated with question-answer pairs. As a result, text-to-speech (TTS) has been used

LREC-COLING 2024, pages 4449–4455
20-25 May, 2024.

to artificially generate audio data from textual QA datasets, e.g., Spoken-SQuAD (Li et al., 2018). While TTS-based approaches have shown potential in facilitating data collection for spoken QA, they rely on a substantial amount of human-generated question-answer pairs.

For data scarcity problem in text-based QA, few-shot and zero-shot models have recently been proposed (Ram et al., 2021; Kuo and Chen, 2022). These models aim to improve QA performance in limited or zero-resource scenarios by using a specific pre-training scheme tailored to the QA task or by transferring cross-lingual knowledge from rich to low resource languages. Unlike these works, our approach addresses the data scarcity problem in spoken QA and it is based on automatically generating question-answer pairs from unannotated data. To the best of our knowledge, data scarcity for spoken QA has not been explored before.

QG models that learn to produce question-answer pairs from input passages have been proposed for domain adaptation (Shakeri et al., 2020; Luo et al., 2022) and data augmentation (Alberti et al., 2019; Puri et al., 2020; Lee et al., 2020) in text-based QA, for zero-shot cross-lingual QA (Shakeri et al., 2021), and for automatically generating a QA dataset from scratch (Ünlü Menevse et al., 2022). These works focus on improving generation performance by training QG models with large amounts of annotated data. In contrast, our approach utilizes limited annotated data both for QG and QA training, and this makes our approach a very realistic scenario for low resource languages and domains.

3. Framework

In this paper, we use the framework shown in Fig. 1 to investigate the effect of automatically generated QA data in limited resource spoken QA settings. The framework consists of QG and QA systems. Both systems are based on pre-trained models. The QG model is fine-tuned using annotated data (paragraphs + question-answer pairs). After fine-tuning, the QG system takes unannotated data (paragraphs only) as input and produces automatically generated question-answer pairs for each input paragraph. The data generated by the QG system are used to fine-tune the spoken QA model together with the annotated data.

The QA system predicts the answers to the questions related to the input paragraphs. We use a cascade QA system (ASR + QA). While training QA model, the generated question-answer pairs are utilized in two different ways. We first use data augmentation where manually annotated data and automatically generated data are merged and shuffled, and the pretrained QA model is fine-tuned on the augmented data. Our preliminary experiments show that this approach has the drawback of relying mostly on the large amounts of generated data, which may contain inaccurate question-answer pairs. Then we propose a two-step fine-tuning approach which allows us to use the generated data effectively together with the limited annotated data. In this fine-tuning approach, the QA model is first fine-tuned on the generated data and then on the manually annotated data. This approach has the advantage of preventing the model to learn from mostly the repetitive or incorrect questions and answers that could result from automatically generated data.

In the framework, we assume that the annotated QA data are limited and there are large amounts of unannotated data. Therefore, the limited annotated data are used to fine-tune both QG and QA models. The pre-trained QG and QA models are multilingual, so this framework can be extended to many low resource languages if there are limited QA data.

4. Experimental Setup

4.1. Question Answering Data

We used SQuAD (Rajpurkar et al., 2016) and Spoken-SQuAD (Li et al., 2018) datasets in the experiments. SQuAD contains 100,000+ question-answer pairs on 536 articles and 23.2K paragraphs totally for train, development and test partitions. Spoken-SQuAD contains ASR transcriptions of the paragraphs in SQuAD and their question-answer pairs. Spoken-SQuAD has a subset of SQuAD question-answer pairs (approximately 37K for train and 5.3K for development partitions) since the pairs were removed from the dataset if the answer to a question did not exist in ASR transcriptions. The development partition of Spoken SQuAD is available in three versions with different levels of added noise: No Noise, Noise V1, Noise V2. Word error rates (WERs) for these sets are 22.73%, 44.22%, and 54.82%, respectively.

4.2. Question-Answer Generation

In our experiments, we first shuffled the articles in the train partition of the QA dataset and then set apart 10% of the articles as our development set to tune the hyperparameters of the models. The remaining articles were split into two equal size disjoint partitions. These partitions are called Part-I and Part-II, respectively. We used the paragraphs and question-answer pairs in Part-I as the annotated data to fine-tune the QG and QA models. To investigate QG and QA performance with limited amounts of annotated data, we further split down the articles in the Part-I partition into 6 varying size subsets. These subsets contain approximately 5%, 10%, 20%, 30%, 40%, and 50% of the articles

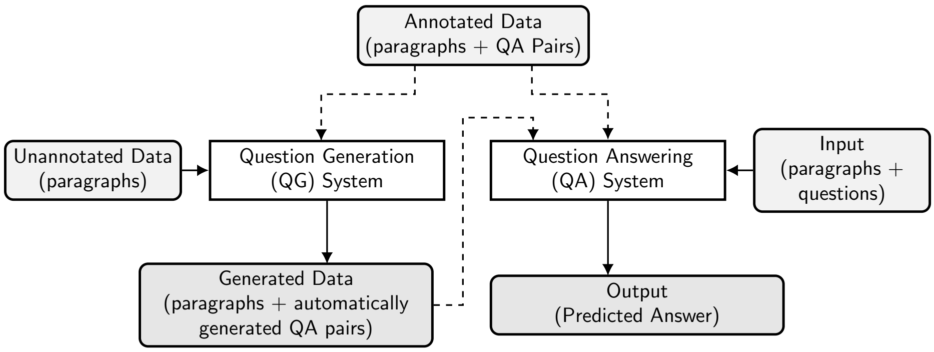

Figure 1: Proposed framework. The dashed lines show the data used in fine-tuning the models. The solid lines show the inputs and outputs.

in the train partition. Note that the larger subsets contain the smaller ones and the largest subset, 50%, corresponds to the whole Part-I data. We treat Part-II as an unannotated partition, and we used only the paragraphs, not the question-answer pairs, in QG to automatically generate question-answer pairs. We trained 6 different QG models, one for each subset coming from Part-I. Then, automatically generated question-answer pairs were produced for the paragraphs in Part-II using these QG models.

For the experiments, we investigated two different setups for QG. In the first setup, we used the SQuAD dataset to produce the automatically generated question-answer pairs. For the generated data, first the SQuAD paragraphs in Part-II partition were replaced with their corresponding ASR transcriptions and these transcriptions were utilized together with the automatically generated question-answer pairs. Then the question-answer pairs where the answer to a question did not exist in the transcriptions were eliminated. A similar approach was utilized while generating Spoken SQuAD from the SQuAD dataset. This setup is referred to as "Eliminated SQuAD" in QG experiments. In the second setup, we used the Spoken-SQuAD dataset to produce the automatically generated question-answer pairs. For the generated data, these pairs were utilized together with the Spoken-SQuAD paragraphs (ASR transcriptions). In this setup, Part-I partition and its subsets have fewer data than those obtained from SQuAD due to using the Spoken-SQuAD dataset. Since Part-II partition only contains the Spoken-SQuAD paragraphs, not the question-answer pairs, it has the same amount of data with the first setup. This setup is referred to as "Only Spoken-SQuAD" in QG experiments.

The QG system is based on mT5 (Xue et al., 2021), an encoder-decoder based transformer network pretrained with multilingual data. The QG system was implemented in Python using the HuggingFace library (Wolf et al., 2020) with the pretrained mT5-Small model. In QG experiments, we used a batch size of 8 with gradient accumulation steps of 32 to achieve an effective batch size of 256. Based on our preliminary experiments on our set apart development set, the learning rate was chosen as 1e-3 and all QG models were fine-tuned for 10 epochs.

4.3. Spoken Question Answering

We performed spoken QA experiments using the Spoken-SQuAD dataset. Similar to QG, we investigated QA performance with limited amount of annotated data. We used the varying sized Spoken-SQuAD data subsets (5%, 10%, 20%, 30%, 40%, and 50%) from Part-I and trained baseline QA models, one for each data subset. The limited annotated data used in QA experiments are the same with the ones used in Part-I partition of the "Only Spoken-SQuAD" setup in the QG experiments. As explained in Section 3, two different fine-tuning strategies, augmentation and 2-step fine-tuning, were employed to integrate the automatically generated data into the QA model. As QA models, we used BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020). The Bert model was English BERT-base-uncased. The QA system was implemented in Python using the HuggingFace library (Wolf et al., 2020). While fine-tuning the QA models, the batch size was set to 12 and the learning rate was set to 3e-5. The number of epochs were determined by evaluating the F1 performance of the models on our set apart development set.

5. Results

5.1. Question-Answer Generation

The QG model was evaluated on dev sets of SQuAD and Spoken SQuAD using BLEU and ROUGE metrics. Table 1 displays the evaluation results of QG models along with the number of generated questions (Nbr Q) from each setup. For Eliminated SQuAD the generated questions were evaluated on SQuAD dev set, while in Only Spoken SQuAD setup spoken SQuAD dev set was used. Increasing the amount of annotated data during training resulted in improvements in both metrics. As expected, Eliminated SQuAD has the lowest number of generated questions due to some examples being removed as a result of ASR errors.

Table 1: Evaluation results and number of question-answer pairs generated from the QG models

Setups	Data Subsets	ROUGE	BLUE 1	BLUE 2	Nbr Q
Eliminated SQuAD	5%	0.305	0.284	0.112	8,537
	10%	0.317	0.295	0.122	7,704
	20%	0.330	0.307	0.133	8,192
	30%	0.340	0.316	0.140	8,458
	40%	0.342	0.320	0.142	7,989
	50%	0.349	0.323	0.146	8,189
Only Spoken SQuAD	5%	0.242	0.222	0.075	24,163
	10%	0.298	0.276	0.108	39,127
	20%	0.310	0.288	0.117	39,024
	30%	0.334	0.309	0.133	39,827
	40%	0.342	0.315	0.139	39,794
	50%	0.345	0.317	0.142	39,931

5.2. Spoken Question Answering

F1 and EM scores of Spoken QA models are given in Fig. 2 (with BERT model) and Fig. 3 (with ELECTRA model), where the horizontal axis shows the amount of the annotated QA data and the vertical axis shows F1 and EM scores. In the experiments, we used the QG models trained with the "Only Spoken-SQuAD" setup and the performance was evaluated on the No Noise dev set (22.73% WER). QA performance of the baseline models trained only on the annotated data are shown with black circles which were then interpolated with a dashed line to point out the general trend in the plots. The scores of the baseline models illustrate that performance improves with increasing annotated data. The models trained only with the automatically generated data show the lowest performance, possibly due to noise in the question-answer pairs. The 2-step fine-tuning approach outperforms both the augmented data approach and the baseline models, showing the importance of the carefully chosen fine-tuning strategy. In more realistic limited data scenarios (i.e., 5%, 10%), performance improvements are more pronounced. 5% subset yields 4.8% relative F1 improvement (from 55.91 to 58.59), 10% subset yields 5.5% relative F1 improvement (from 60.79 to 64.16) with BERT model. With larger data (i.e., 40%, 50%), on the other hand, the gap between the solid and dashed lines gets smaller. Similar trend in performance is observed with ELECTRA based QA models. For instance, 5% subset results in a 3.7% relative F1 improvement (from 66.57 to 69.09), 10% subset results in a 5.5% relative

Table 2: F1 scores of QA models on Spoken SQuAD dev sets with different noise levels.

	No Noise	Noise V1	Noise V2
Generated Data	51.43	41.35	30.77
Annotated Data	60.79	45.58	34.46
2-Step	64.16	48.01	36.42

F1 improvement (from 70.62 to 74.55). Note that fine-tuning the QA model with the entire Spoken-SQuAD training dataset (100%) results in F1/EM scores of 74.20/64.02 with BERT and 80.69/72.01 with ELECTRA. When using 5% and 50% annotated data subsets for the BERT model, the F1/EM scores are 55.91/43.86 and 70.16/59.11, respectively. For the ELECTRA model, the corresponding F1/EM scores are 66.57/55.88 and 77.39/67.95.

Assessing the impact of noisy question-answer pairs generated by QG on QA model performance directly is a challenging task. Determining the degree of noise within the generated data necessitates manual evaluation, which is resource-intensive. Therefore, our evaluation primarily relied on extrinsic measures, where the noisy data were evaluated within the context of downstream QA tasks. Our findings indicate that augmenting annotated data for QG training enhances QA performance (refer to Fig.2 and Fig.3). Notably, this improvement extends to models trained solely on automatically generated data, possibly due to the reduction in noisy pairs with annotated data utilization during QG training.

We also compared the F1 performance of spoken QA models trained with annotated data (baseline) and 2-step fine-tuning approach using the question-answer pairs obtained with the "Eliminated SQuAD" setup. We observed a similar trend in both models trained with the 2-step fine-tuning approach, resulting in better performances than the baseline model. Generating question-answer pairs from noisy ASR transcriptions ("Only Spoken-SQuAD") is still competitive with using clean data during question-answer generation ("Eliminated SQuAD") as there is no significant performance difference between two settings.

We conducted additional experiments with BERT QA model using noisy Spoken SQuAD dev sets. Based on the best improvement over the baseline from the previous experiments, we utilized the QA model fine-tuned on 10% subset of Spoken SQuAD. Table 2 presents the performance of these models on No Noise (22.73% WER), Noise V1 (44.22% WER), and Noise V2 (54.82% WER) dev sets. The No Noise results match those of the Spoken-SQuAD results in Fig. 2. The models tested on the Noise V2 dev set had the lowest scores due to the high noise level. The results indicate that incorporating automatically generated data using a

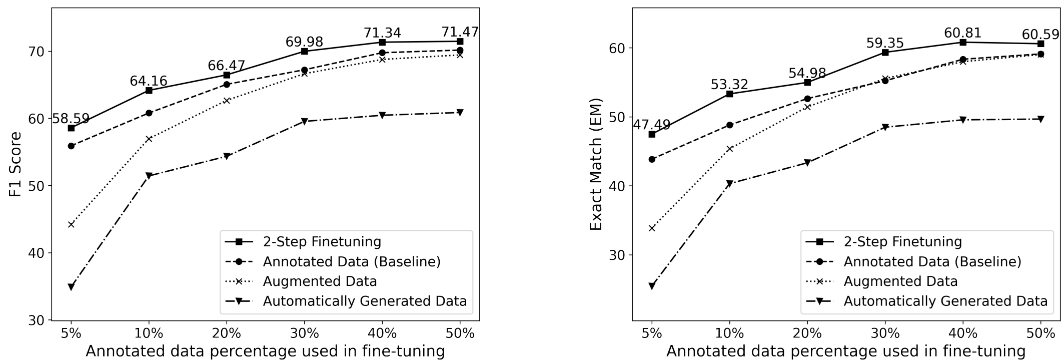

Figure 2: F1 and EM results with BERT trained with Spoken SQuAD 5-50% settings

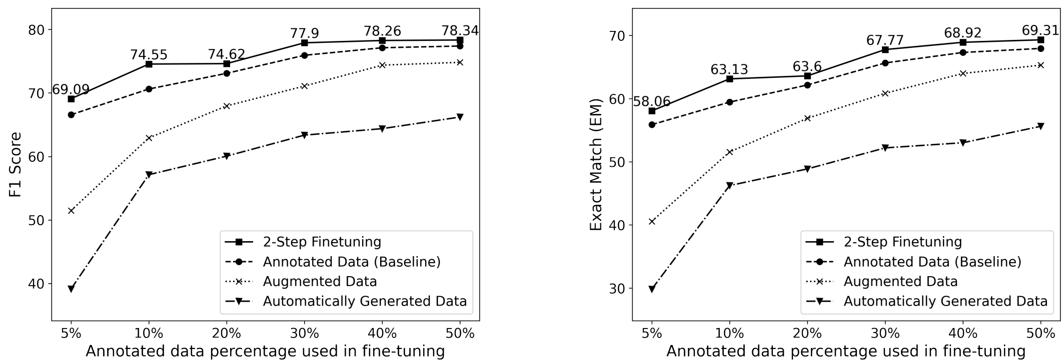

Figure 3: F1 and EM results with ELECTRA trained with Spoken SQuAD 5-50% settings

2-step fine-tuning approach improves performance across all sets. The relative gains in F1 score are consistent across all models, at around 5% (from 60.79 to 64.16 for No Noise, from 45.58 to 48.01 for Noise V1, and from 34.46 to 36.42 for Noise V2).

6. Conclusion

In this paper, we addressed the problem of data scarcity in spoken QA and proposed a framework to improve the performance of spoken QA models. Our framework is based on careful selection of fine-tuning strategy for QA and uses limited annotated data for question generation (QG) and spoken QA training. Spoken QA may suffer more from data scarcity than textual QA due to the high cost of data collection. To the best of our knowledge, data scarcity for spoken QA has not been explored before. Unlike previous works on Spoken QA that use large amounts of annotated data for training, our approach leverages limited annotated data for both QG and spoken QA training, which is a more realistic scenario for low-resource languages and domains. We showed that the proposed framework is effective for limited data spoken QA.

7. Ethical statement

The data utilized in this research is randomly sourced from publicly available SQuAD and Spoken SQUAD datasets to fine-tune the publicly available pretrained QG and QA models. Both datasets do not contain any details or identifiers that could be used to personally identify individuals, including information such as names, addresses, phone numbers, email addresses, social security numbers, or any other data. With the assumption of the impartiality of pretrained models, we do not expect a potential bias from the QG and QA models. The randomization of the subsets used to finetune question generation models aimed to include a diverse dataset while mitigating bias towards any particular topic. To facilitate reproducibility, we plan to publicly release the generated dataset alongside the implementation details.

8. References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proc. ACL*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Yung-Sung Chuang, Chi-Liang Liu, Hung-Yi Lee, and Lin-shan Lee. 2019. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. In *INTERSPEECH*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proc. ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, pages 4171–4186, Minneapolis, Minnesota.

Chia-Chih Kuo and Kuan-Yu Chen. 2022. Toward zero-shot and zero-resource multilingual question answering. In *IEEE Access*, volume 10, pages 99754–99761.

Chia-Hsuan Lee, Yun-Nung Chen, and Hung-Yi Lee. 2019a. Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7300–7304.

Chia-Hsuan Lee, Hung-yi Lee, Szu-Lin Wu, Chi-Liang Liu, Wei Fang, Juei-Yang Hsu, and Bo-Hsiang Tseng. 2019b. Machine comprehension of spoken content: Toefl listening test and spoken squad. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(9):1469–1480.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proc. ACL*, pages 208–224, Online. Association for Computational Linguistics.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. In *INTERSPEECH*.

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu-wen Yang, Hsuan-Jui Chen, Shuyan Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, and Lin-shan Lee. 2022. DUAL: Discrete spoken unit adaptive learning for textless spoken question answering. In *Proc. Interspeech*, Korea. ISCA.

Hongyin Luo, Shang-Wen Li, Mingye Gao, Seunghak Yu, and James Glass. 2022. Cooperative self-training of machine reading comprehension. In *Proc. NAACL*, pages 244–257, Seattle, United States. Association for Computational Linguistics.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proc. EMNLP*, pages 5811–5826, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. EMNLP*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. In *Proc. ACL*, pages 3066–3079, Online. ACL.

Siamak Shakeri, Noah Constant, Mihir Kale, and Linting Xue. 2021. Towards zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 35–45, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proc. EMNLP*, pages 5445–5460, Online. Association for Computational Linguistics.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. *Proc. INTERSPEECH*.

Merve Ünlü and Ebru Arisoy. 2021. Uncertainty-aware representations for spoken question answering. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 943–949.

Merve Ünlü, Ebru Arisoy, and Murat Saraçlar. 2019. Question answering for spoken lecture processing. In *Proc. ICASSP*, pages 7365–7369. IEEE.

Merve Ünlü Menevse, Yusufcan Manav, Ebru Arisoy, and Arzucan Özgür. 2022. A framework for automatic generation of spoken question-answering data. In *Findings of EMNLP 2022*, Abu Dhabi. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proc. EMNLP*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. NAACL*, pages 483–498, Online. Association for Computational Linguistics.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021. Knowledge distillation for improved accuracy in spoken question answering. In *Proc. ICASSP*, pages 7793–7797. IEEE.