IMPROVING SPOKEN QUESTION ANSWERING USING CONTEXTUALIZED WORD REPRESENTATION

Dan Su and *Pascale Fung*

Department of Electronic and Computer Engineering
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology, Clear Water Bay
dsu@connect.ust.hk, pascale@ece.ust.hk

ABSTRACT

While question answering (QA) systems have witnessed great breakthroughs in reading comprehension (RC) tasks, spoken question answering (SQA) is still a much less investigated area. Previous work shows that existing SQA systems are limited by catastrophic impact of automatic speech recognition (ASR) errors [1] and the lack of large-scale real SQA datasets [2]. In this paper, we propose using contextualized word representations to mitigate the effects of ASR errors and pretraining on existing textual QA datasets to mitigate the data scarcity issue. New state-of-the-art results have been achieved using contextualized word representations on both the artificially synthesised and real SQA benchmark data sets, with 21.5 EM/18.96 F1 score improvement over the sub-word unit based baseline on the Spoken-SQuAD [1] data, and 13.11 EM/10.99 F1 score improvement on the ODSQA data [2]. By further fine-tuning pre-trained models with existing large scaled textual QA data, we obtained 38.12 EM/34.1 F1 improvement over the baseline of fine-tuned only on small sized real SQA data.

Index Terms— contextual word representation, spoken question answering, supervised pre-training, BERT

1. INTRODUCTION

With the recent studies on reading comprehension by machine, we have witnessed several breakthroughs in question answering (QA) systems [3]. However, machine comprehension of spoken content, which is much more difficult and time-consuming than plain text content for humans, is a much less investigated area. Spoken question answering (SQA) is the task to automatically find relevant answers to given textual or spoken questions from spoken documents, such as online lecture videos [4], listening comprehension exams [5], or movie clips [6].

Typical SQA systems including two parts. They first transcribe spoken content into text by automatic speech recognition (ASR) engine, and then information retrieval (IR) [7] or question answering techniques [8] will be used to find the relevant answers from the ASR hypothesis. Previous work [9] has shown the catastrophic impact of the ASR errors in degrading SQA system performances, and related investigations have been conducted to address the problem. [2, 10] proposed using phonetic sub-word units as word representations to mitigate the impact of speech recognition errors, based on the assumption that some sub word units may be correctly transcribed even if the transcribed word itself is wrong. [9] proposed using adversarial domain adaptation methods, trying to adapt the source domain data (reference transcriptions) to the target domain (ASR hypothesis), to improve the robustness to ASR errors. However, their methods requires large-scale source domain data (reference transcriptions of the spoken documents), which is too costly to obtain in real SQA systems.

On the other hand, human are more robust to ASR errors in understanding language, because we use contextual information, common sense and reasoning ability for automatically error words correction. Thus intuitively, effective contextual representation of words in a SQA system will help mitigate the catastrophic effects of ASR errors. Recently, deep bidirectional language models pre-trained on large amount of unlabeled text by jointly condition on both left and right contexts, have shown its power in representing contextual information. Breakthrough performances on many language understanding tasks such as the GLUE benchmark and the SQuAD QA task [11, 3] have been obtained by adding an extra layer and fine-tuning on task-specic supervised data on the pre-trained model. Nevertheless, the capability of pre-trained contextualized word representations in mitigating the effect of ASR errors in SQA has not been explored yet.

Further more, another issue which hinders the performance of real SQA system is the lack of real and large scale SQA data sets. Existing large scale SQA data like Spoken-SQuAD [1] is artificially generated using Text-to-speech (TTS) and then transcribed back by ASR on existing QA data set SQuAD [12]. As [9] has shown that the performance drops a lot when the training and testing data mismatch, it

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

is still one step away from real SQA when training on synthesised data. On the other hand, creating real SQA data set is very expensive and the existing small data size is far away from enough to train a neural SQA model which usually has millions of parameters. The largest real Chinese SQA dataset ODSQA [2] only has around three thousand QA pairs. Considering there are a lot of QA data available, we thus investigate transfer learning, by fine-tuning models pre-trained with existing large-scale QA data sets first, to boost the real SQA system performance.

In this paper, we proposed (1) using contextualized word representations (from BERT) to mitigate the effects of ASR error (2) using supervised pre-training with large amounts of existing QA data sets, to help improve the fine-tuning performance of real SQA data in SQA system. State-of-the-art results have been obtained on two SQA benchmark datasets.

2. SPOKEN QUESTION ANSWERING

In this task, when the machine is given a spoken document, it needs to find the answer of a question from the spoken document. SQA can be solved by the concatenation of an ASR module and a question answering module. Given the ASR hypotheses of a spoken document and a question, the question answering module can output a text answer.

We compute the Exact Match (EM) and Macro-averaged F1 scores (F1) between the predicted text answer and the ground-truth answer to evaluate the model accuracy, which are the standard evaluation metrics used in the SQuAD QA task [12].

3. METHODOLOGY

The architecture of contextualized word representation based SQA system is shown in Fig. 1. The input X, which is a word sequence is first represented as a sequence of embedding vectors, one for each word, in $l1$. Then the BERT Transformer encoder captures the contextual information for each word via self-attention, generating contextual embeddings in $l2$. Then the semantic representation of the input QA pairs will be forward to a MLP layer, in our case we use linear NN, to predating the starting and end position of the answer spans.

Lexicon Encoder ($l1$): The input $X = x_1, ..., x_m$ is a sequence fo tokens of length m. Following [3], the first token x_1 is always the [CLS] token. and we seperate the question and document with a special token [SEP]. The lexicon encoder maps X into a sequecen of input embedding vectors, by summing the corresponding word, segment and positional embeddings for each token.

Transformer Encoder (Contextual embedding encoder) ($l2$): We use the multi layer bidirectional Transformer encoder BERT [3], to map input representation vectors ($l1$) into a sequence of contextual embedding vectors ($l2$). Both the $BERT_{BASE}$ and $BERT_{LARGE}$ models have been adopted in our experiments.

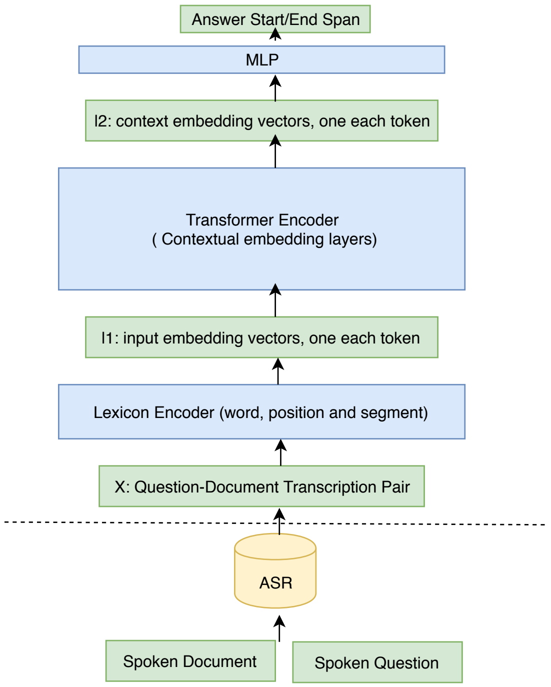
Answer Start/End Span
MLP
l2: context embedding vectors, one each token
Transformer Encoder
(Contextual embedding layers)
l1: input embedding vectors, one each token
Lexicon Encoder (word, position and segment)
X: Question-Document Transcription Pair
ASR
Spoken Document
Spoken Question

Fig. 1. Architecture of our Contextualized Word Representation (BERT) - based SQA System

Output Layer (MLP): We use linear NN follwed by softmax function to generate the starting and end position probability vectors. Two new parameters for the fine tuning: a start vector $S \in R^H$ and end vector $E \in R^H$. Let's denote the context embedding vector for the ith input token is $Ti \in R^H$. We can calculate the probability of word i being the start of the answer span as a dot product between T_i and S followed by a softmax over all of the words in the documents. And the training objective is the log-likelihood of the correct start and end position.

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \tag{1}$$

Multi-Dataset Pre-training: Empirically, the data feeding order when pre-training or fine-tuning will effect the performance of the model. Thus in terms of the pre-training procedure with multiple training sets, we propose **multi-datasets** for data feeding. The method follows the idea of **multi-task** learning. We consider different datasets as different QA tasks and leverage the model to fully explore the general semantic representations of the samples in the training datasets. During multi-dataset pre-training, we combine all the training datasets and shuffle them to reduce the reliance on the order of the data as shown in Fig 2.

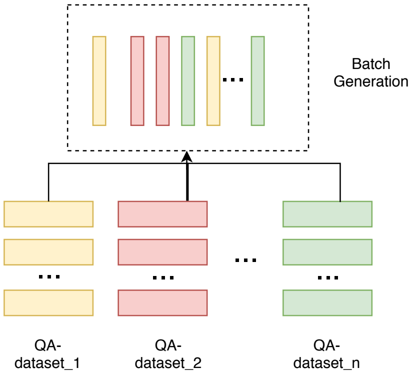

Fig. 2. Multi-dataset pre-training batch generation

4. EXPERIMENTS

4.1. SQA Datasets

Spoken-SQuAD: Spoken-SQuAD [1] is an automatically generated corpus by applying Google Text-to-speech (TTS) and then transcribed back by ASR on SQuAD [12], the document is in spoken form and the question is in text form. There are 37,111 and 5,351 question answer pairs in the training and testing sets respectively, and the word error rate (WER) of both sets is around 22.7%.

ODSQA: Open Domain Spoken Question Answering (ODSQA) [2] is the largest real chinese SQA dataset for extraction-based QA task. The basic data statistics are shown in row 1 in Table 2. Both the document and the question are in spoken form and the text-formed answer to each question is always a span in the document. The dataset was created by recruiting native Chinese speakers to read the questions and paragraphs in the development set of DRCD[13], and the WER is around 19%.

4.2. QA Datasets for Pre-training

The multiple QA datasets used in pre-training the BERT-based SQA system are shown in Table 1. All of them are large scaled and publicly available. The detailed descriptions for each of them can be found in [14, 15, 16, 17, 18].

While [1] reported that training on transcriptions with ASR errors are better than training on text, two artificially generated SQA data sets [2] are investigated for pre-training. DRCD-TTS are generated by applying the previous mentioned TTS-ASR pipeline procedures on the DRCD dataset[13], and DRCD-backtrans are generated using back translation with English as pivot language on the DRCD data set. The detailed statistics of the data are shown in Table 2.

Dataset	Source	Question	Size
NewsQA	News	Crowd	120K
TriviaQA	Snippets	Trivia	95K
SearchQA	Snippets	Trivia	140K
HotpotQA	Wikipedia	Crowd	113K
NaturalQuestions	Wikipedia	Query	30K

Table 1. Characterization of different QA datasets for pre-training Spoken-SQuAD based SQA system

Dataset	QA	WER-D	Avg	Avg
	-pairs	(%)	D Len	Q Len
ODSQA	1,465	19.11	428	22
DRCD-TTS	16746	33.63	332	20
DRCD-backtrans	15238	45.64	439	20
DRCD	26936	0	436	21

Table 2. Data statistics of DRCD-TTS, DRCD-backtrans and DRCD for pre-training ODSQA-based real SQA system

4.3. Training Details

Our implementation is based on the Pytorch implementation of Transformers[1]. We use Adamax as our optimizer with a learning rate of 3e-5. The maximum training epochs was set to 2. For the Spoken-SQuAD related experiments we use a batch size of 12 for $BERT_{BASE}$ and 2 for $BERT_{LARGE}$, and the maximum total input sequence length (including both the question and documents) is set to 384 and document stride of 128. For the ODSQA involved experiments, the batch size is 6 with *bert-base-chinese*, and maximum total input sequence length 512 with document stride of 256.

4.4. Results

Mitigating ASR errors by Contextualized Word Representations

The results of the experiments using BERT-based pre-trained contextualized word representations in SQA system, fine-tuned with the SQA data, is shown in Table 3 and Table 4. We also include the results on Spoken-SQuAD from [9], which propose using adversial domain adaptation method to mitigating ASR error. Compared with the baseline systems using BiDAF [8] with various combinations of sub-word unit embeddings input representations, and on differen training data set combinations, absolute improvements on EM and F1 score have been obtained on both datasets by using contextualize word representations from BERT.

Supervised Pre-training Improve SQA Performance on Small Datasets

For the lack of large scale SQA dataset to train real SQA system, we further investigate whether and how the supervised

[1] https://github.com/huggingface/transformers/

models	EM	F1
WORD	44.34	57.37
WORD+CHAR	44.45	57.6
WORD+PHONEME	45.58	58.25
WORD+SYLLABLE	45.61	58.25
WORD+CHAR+PHONEME+SYLLABLE	**45.78**	**58.71**
WORD (+ GAN) [9]	**51.10**	**63.11**
$BERT_{base}$	**59.71**	**70.94**
$BERT_{large}$	**67.37**	**77.67**

Table 3. Performance comparison of $BERT_{base}$ and $BERT_{large}$ contextual representations VS various embedding baselines (based on BiDAF[8]) over Spoken SQuAD testing set

models	EM	F1
WORD + PINYIN (+ DRCD)	55.49	68.79
WORD + PINYIN (+ DRCD-TTS)	51.74	64.59
WORD + PINYIN (+ DRCD-back)	48.2	62.82
WORD + PINYIN (+ DRCD+TTS+backtrans)	**59.52**	**70.95**
$BERT_{chinese}$ (+ DRCD)	**69.97**	**79.61**
$BERT_{chinese}$ (+ DRCD-TTS)	**67.78**	**77.87**
$BERT_{chinese}$ (+ DRCD-back)	**66.48**	**76.38**
$BERT_{chinese}$ (+ DRCD+TTS+backtrans)	**72.63**	**81.94**

Table 4. Performance comparison of *bert-base-chinese* contextual representations VS sub-word unit representations (based on BiDAF) on different SQA training data combinations over the ODSQA dataset

pre-training with existing large-scale QA data sets will help improve SQA system performance.

We conducted experiments by first pre-train the BERT-based model using the existing QA data sets as described in Section 4.2, and then fine-tune on the target domain SQA datasets. The pre-training and fine-tuning are identical and will stop if no improvement can be obtained on the corresponding development set.

As we can see from Table 5, the pre-training using multiple data sets seems did not improve the performance on the Spoken-SQuAD test data, and the performance even decrease when increasing the size of pre-training datasets. One possible explanation for this might be because of the domain mismatch between the pre-training QA data sets and the target Spoken-SQuAD data. Pre-training with more large-scaled, ASR-error clean QA datasets, will only fine-tune the general contextualized word representations be more domain specific and insensitive to SQA data which have ASR errors. Thus supervised pre-training not always helps. While [1] reported that training on transcriptions with ASR errors are better than training on text for SQA system, we further conducted the pre-training experiments using multiple datasets with ASR errors, which are more similar to the SQA data. The data statistics including the Word Error Rate (WER) are shown in Table 2. We split the ODSQA data set into training and testing sets, which contain 1238 and 223 examples respectively. The performances was evaluated on ODSQA test data, before and after further fine-tuning with the ODSQA training data, on different pre-training models. From the comparison results on the ODSQA test data (row (a) vs row (i)), we can see that with small sized real SQA data, which is inadequate to train a good SQA model, we can obtain satisfied performane by pre-training on other similar synthesised SQA datasets. The more similar dataset we have in pre-training, the more performance gain will be obtained (row (h) and row (i)). The results also show that with pre-training using large-scaled artificially generated SQA data, fine-tune using small sized real SQA dataset still further improves SQA performance.

models	EM	F1
$BERT_{large}$	**67.37**	**77.67**
$BERT_{large}$ (+ Multi-20K)	67.74	77.59
$BERT_{large}$ (+ Multi-75K)	65.28	75.20
$BERT_{large}$ (+ Multi-375K)	65.24	74.97

Table 5. Spoken-SQuAD fine-tuning on models pre-trained on different sized multiple QA dataset

models		EM	F1
$BERT_{chinese}$	(a)	34.53	47.47
$BERT_{chinese}$ (+ DRCD)	(b)	67.26	77.37
+ fine-tune	(c)	71.75	81.13
$BERT_{chinese}$ (+ DRCD-TTS)	(d)	64.57	75.60
+ fine-tune	(e)	64.13	76.11
$BERT_{chinese}$ (+ DRCD-backtrans)	(f)	64.57	73.80
+ fine-tune	(g)	67.71	76.28
$BERT_{chinese}$ (+ DRCD+TTS+backtrans)	(h)	69.06	79.90
+ fine-tune	(i)	**72.65**	**81.57**

Table 6. Performance comparison over the ODSQA test set, w/o ODSQA training set fine-tuning on pre-trained models by different SQA dataset

5. CONCLUSION

In this paper, we proposed using contextualized word representations to mitigate the effects of ASR error on SQA system, state-of-the-art results have been achieved on two SQA benchmark data sets, with 21.59/18.96 EM/F1 score improvement over the sub word unit based baseline on Spoken-SQuAD test set, and 13.11/10.99 EM/F1 score improvement on ODSQA data set. We also investigate transfer learning by fine-tuning models pre-trained with existing large scaled QA data. We obtained 38.12/34.1 EM/F1 score improvement over the baseline which is only fine-tuned on small sized real SQA data ODSQA.

6. REFERENCES

[1] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee, "Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension," *arXiv preprint arXiv:1804.00320*, 2018.

[2] Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee, "Odsqa: Open-domain spoken question answering dataset," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 949–956.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[4] Merve Ünlü, Ebru Arisoy, and Murat Saraçlar, "Question answering for spoken lecture processing," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7365–7369.

[5] Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee, "Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine," *arXiv preprint arXiv:1608.06378*, 2016.

[6] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4631–4640.

[7] Sz-Rung Shiang, Hung-yi Lee, and Lin-shan Lee, "Spoken question answering using tree-structured conditional random fields and two-layer random walk," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi, "Bidirectional attention flow for machine comprehension," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[9] Chia-Hsuan Lee, Yun-Nung Chen, and Hung-Yi Lee, "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)*. IEEE, 2019, pp. 7300–7304.

[10] Chia-Hsuan Lee, Hung-Yi Lee, Szu-Lin Wu, Chi-Liang Liu, Wei Fang, Juei-Yang Hsu, and Bo-Hsiang Tseng, "Machine comprehension of spoken content: Toefl listening test and spoken squad," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[11] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[12] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[13] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai, "Drcd: a chinese machine reading comprehension dataset," *arXiv preprint arXiv:1806.00920*, 2018.

[14] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman, "Newsqa: A machine comprehension dataset," *arXiv preprint arXiv:1611.09830*, 2016.

[15] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551*, 2017.

[16] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho, "Searchqa: A new q&a dataset augmented with context from a search engine," *arXiv preprint arXiv:1704.05179*, 2017.

[17] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018.

[18] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov, "Natural questions: a benchmark for question answering research," *Transactions of the Association of Computational Linguistics*, 2019.