ODSQA: OPEN-DOMAIN SPOKEN QUESTION ANSWERING DATASET

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, Hung-Yi Lee

College of Electrical Engineering and Computer Science, National Taiwan University, Taiwan

ABSTRACT

Reading comprehension by machine has been widely studied, but machine comprehension of spoken content is still a less investigated problem. In this paper, we release Open-Domain Spoken Question Answering Dataset (ODSQA) with more than three thousand questions. To the best of our knowledge, this is the largest real SQA dataset. On this dataset, we found that ASR errors have catastrophic impact on SQA. To mitigate the effect of ASR errors, subword units are involved, which brings consistent improvements over all the models. We further found that data augmentation on text-based QA training examples can improve SQA.

Index Terms— spoken question answering

1. INTRODUCTION

Machine comprehension and question answering on text have significant progress in the recent years. One of the most representative corpora is the Stanford Question Answering Dataset (SQuAD) [1], on which deep neural network- (DNN-) based models are comparable with human. The achievements of the state-of-the-art question answering models demonstrate that machine has already acquired complex reasoning ability. On the other hand, accessing large collections of multimedia or spoken content is much more difficult and time-consuming than plain text content for humans. It is therefore highly attractive to develop Spoken Question Answering (SQA) [2, 3, 4, 5], which requires machine to find the answer from spoken content given a question in either text or spoken form.

In SQA, after transcribing spoken content into text by automatic speech recognition (ASR), typical approaches use information retrieval (IR) techniques [6] or knowledge bases [7] to find the proper answer from the transcriptions. Another attempt towards machine comprehension of spoken content is TOEFL listening comprehension by machine [8]. TOEFL is an English examination that tests the knowledge and skills of academic English for English learners whose native languages are not English. Deep-based models including attention-based RNN[8] and tree-structured RNN[9] were used to answer TOEFL listening comprehension test. Transfer learning for Question Answering (QA) is also studied on this task[10]. However, TOEFL listening comprehension test is a multi-select question answering corpus, and its scale is not large enough to support the training of powerful listening comprehension models. Another spoken question answering corpus is Spoken-SQuAD[11], which is generated from SQuAD dataset through Google Text-to-Speech (TTS) system. The spoken content is then transcribed by CMU sphinx[12]. Several state-of-the-art question answering models are evaluated on this dataset, and ASR errors seriously degrade the performance of these models. On Spoken-SQuAD, it has been verified that using sub-word units in SQA can mitigate the impact of ASR errors. Although Spoken-SQuAD is large enough to train state-of-the-art QA models, it is artificially generated, so it is still one step away from real SQA.

To further push the boundary of SQA, in this paper, we release a large scale SQA dataset – Open-Domain Spoken Question Answering Dataset (ODSQA). The contribution of our work are four-fold:

- First of all, we release an SQA dataset, ODSQA, with more than three thousand questions. ODSQA is a Chinese dataset, and to the best of our knowledge, the largest real[1] SQA dataset for extraction-based QA task.
- Secondly, we found ASR errors have catastrophic impact on real SQA. We tested numbers of state-of-the-art SQuAD models on ODSQA, and reported their degrading performance on ASR transcriptions.
- Thirdly, we apply sub-word units in SQA to mitigate the impact of speech recognition errors, and this approach brings consistent improvements experimentally.
- Last but not the least, we found that back-translation, which has been applied on text QA [13] to improve the performance of models, also improve the SQA models.

2. RELATED WORK

Most QA work focuses on understanding text documents[14, 15, 1, 16]. The QA task has been extended from text to images [17, 18, 19, 20] or video descriptions [21, 22, 23]. In

Thanks to Delta Research Center and Delta Electronics, Inc. for collecting the DRCD dataset.

[1] not generated by TTS as Spoken-SQA

the MovieQA task[24], the machine answers questions about movies using video clips, plots, subtitles, scripts, and DVS. Usually only text information (e.g., the movie's plot) is considered in the MovieQA task; learning to answer questions using video is still difficult. Machine comprehension of spoken content is still a less investigated problem.

To mitigate the impact of speech recognition errors, we use sub-word units to represent the transcriptions of spoken content in this paper. Using sub-word unit is a popular approach for speech-related down-stream task and has been applied to spoken document retrieval[25], spoken term detection [26][27], spoken document categorization[28], and speech recognition[29]. It has been verified that sub-word units can improve the performance of SQA [11] . However, the previous experiments only conducted on an artificial dataset. In addition, the previous work focuses on English SQA, whereas we focus on Chinese SQA in this paper. There is a big difference between the subword units of English and Chinese.

To improve the robustness to speech recognition errors, we used back-translation as a data augmentation approach in this paper. Back-translation allows the model to learn from more diversified data through paraphrasing. Back-translation was also studied in spoken language understanding and text-based QA as a data augmentation approach. In cross lingual spoken language understanding, training with the back-translation data via target language will make the model adaptive to translation errors [30][31]. In text-based QA, back-translation was used to paraphrase questions[32] and paraphrase documents[13].

3. TASK DESCRIPTION

3.1. Data Format

In this paper, we introduce a new listening comprehension corpus, Open-Domain Spoken Questions Answering Dataset (ODSQA). Each example in this dataset is a triple, (q, d, a). q is a question, which has both text and spoken forms. d is a multi-sentence spoken-form document. The answer a is in text from, which is a word span from the reference transcription of d. An overview architecture of this task is shown in figure 1.

3.2. Data Collection

To build a spoken version QA dataset, we conducted the following procedures to generate spoken documents and questions. Our reference texts are from Delta Reading Comprehension Dataset (DRCD), which is an open domain traditional Chinese machine reading comprehension (MRC) dataset [33]. Each data example in DRCD is a triple of (q, d, a) in which q is a text-form question, d is a multi-sentence text-form document that contains the answer a as an extraction segment. In DRCD, training set contains 26,936 questions with 8,014 paragraphs, the development set contains 3,524 questions with 1,000 paragraphs and the testing set contains 3,485 questions with 1,000 paragraphs. The training set and development set are publicly available, while the testing set is not. Therefore, the DRCD training set was used as the reference text of the training set of ODSQA, and the DRCD development set was used as the reference text of the testing set of ODSQA.

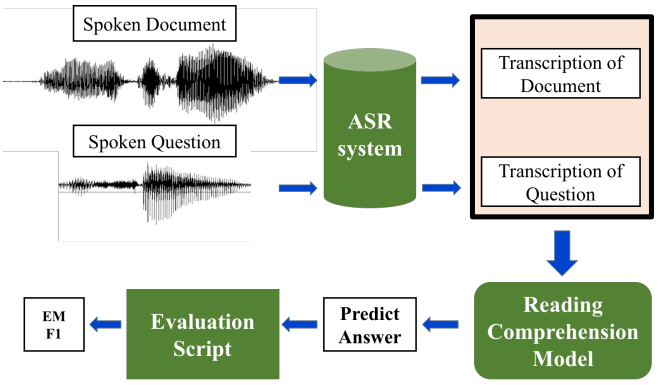

Fig. 1. Flow diagram of the SQA system and the standard evaluation method. Given a spoken document and a spoken or text question, an SQA system, which is a concatenation of ASR module and reading comprehension module, can return a predicted text answer. This predicted answer is a span in the ASR transcription of the spoken document and will be evaluated by EM/F1 scores.

20 speakers were recruited to read the questions and paragraphs in the development set of DRCD. All the recruited speakers were native Chinese speakers and used Chinese as their primary language. For document, each sentence was shown to speaker respectively. The speaker was required to speak one sentence at a time. All the sentences of the same document were guaranteed to be spoken by the same speaker. Because in the real-life user scenario, it is more possible that an user enters a spoken question, and machine answers the question based on an already recorded spoken document collection. The document and the question from the same data example do not have to be recorded by the same speakers.

We collected 3,654 question answer pairs as the testing set. The corpus is released[2]. The speech was all sampled at 16 kHz due to its common usage among the speech community, but also because the ASR model we adopted was trained on 16 kHz audio waveforms. An example of a corresponding pair between DRCD and ODSQA is shown in column(1) and (2) of Table 1. The detailed information of ODSQA about the speakers, audio total length and Word Error Rate are listed in row(1) of Table 2.

[2]ODSQA: OPEN-DOMAIN SPOKEN QUESTION ANSWERING DATASET
https://github.com/chiahsuan156/ODSQA

3.3. Evaluation Metrics

In this task, when the model is given a spoken document, it needs to find the answer of a question from the transcriptions of the spoken document. SQA can be solved by the concatenation of ASR module and reading comprehension module. Given a query and the ASR transcriptions of a spoken document, the reading comprehension module can output a text answer. The most intuitive way to evaluate the text answer is to directly compute the **Exact Match (EM)** and **Macro-averaged F1 scores(F1)** between the predicted text answer and the ground-truth text answer. If the predicted text answer and the ground-truth text answer are exactly the same, then the EM score is 1, otherwise 0. The F1 score is based on the precision and recall. Precision is the percentage of Chinese characters in the predicted answer existing in the ground-truth answer, while recall is the percentage of Chinese characters in the ground-truth answer also appearing in the predicted answer. The EM and F1 scores of each testing example are averaged to be the final EM and F1 score. We used the standard evaluation script from SQuAD[1] to evaluate the performance.

4. PROPOSED APPROACH

ASR errors are inevitable, and they can hinder the reasoning of QA models. However, when a transcribed word is wrong, some phonetic sub-word units in the word may still be correctly transcribed. Therefore, building word representation from sub-word units may mitigate the impact of ASR errors. Pingyin-token sequence of words are used in our experiments. Pinyin, literally meaning "spell out the sound", is the Romanized phonetic transcription of the Chinese language. Each Chinese character consists of one pingyin syllable, and one syllable is composed of a number of pingyin-tokens. We adopt one-dimensional Convolution Neural Network (1-D CNN) to generate the word representation from the pingyin-token sequence of a word, and this network is called Pingyin-CNN. Our proposed approach is the reminiscent of Char-CNN [34, 35], which apply 1-D CNN on character sequence to generate distributed representation of word for text classification task. Pingyin-CNN is illustrated in Figure 2. We explain how we obtain feature for one word with one filter. Suppose that a word W consists of a sequence of pingyin-tokens $P = [p_1, ..., p_l]$, where l is the number of pingyin-tokens of this word. Let $H \in R^{C \times d}$ be the lookup table pingyin-token embedding matrix, where C is the number of pingyin-tokens, and d is the dimension of the token embedding. In other words, each token corresponds to a d-dimensional vector. Given P, after looking up table, the intermediate token embedding $E \in R^{l \times d}$ is obtained. The convolution between E and a filter $F \in R^{k \times d}$ is performed with stride 1 to obtain one-dimension vector $Z \in R^{l-k+1}$. After

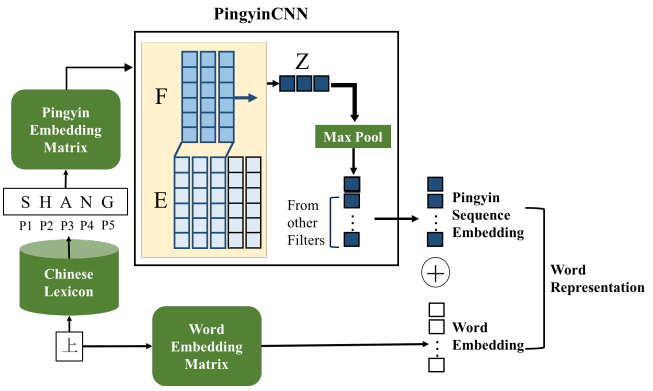

Fig. 2. Illustration of enhanced word embedding. For a given input word W at the bottom, a sequence of pingyin-tokens $P = [p_1, ..., p_l]$ are obtained by looking up in the Chinese lexicon. Each pingyin-token is mapped to a vector R^d and concatenated to form intermediate matrix E. E is fed into the 1-D convolutional module. The output Z is further fed into max-pooling layer, and a scalar value is generated. All the scalars from various filters will form the phoneme sequence embedding. Then the pingyin sequence embedding is further concatenated with word embedding as the input of reading comprehension model. In this illustration, the Chinese word 上 (means "up" in English) consists of five pingyin-tokens.

max pooling over Z, we obtain a scalar value. With a set of filters, with the above process, we obtain a pingyin-token sequence embedding. The size of the pingyin-token sequence embedding is the number of filters. The filter is essentially scanning pingyin-token n-gram, where the size of n-gram is the height of the filter (the number of k above). The pingyin-token sequence embedding is concatenated with typical word embedding to obtain new word representation as the input of reading comprehension model. All the parameters of filters and pingyin-token embedding matrix H are end-to-end learned with reading comprehension model.

5. EXPERIMENTS

5.1. Experimental Setup

- **Speech Recognition**: We used the iFLYTEK ASR system [3] to transcribe both the spoken document and spoken question.
- **Pre-processing**: We used jieba-TW[4], a python library specialized for traditional Chinese, to segment sentence into words. The resulting word vocabulary size for

[3] iFLYTEK ASR system
https://www.xfyun.cn/doccenter/asr
[4] jieba-zh-TW:
https://github.com/ldkrsi/jieba-zh-TW

Table 1. An example in ODSQA and the corresponding reference texts in DRCD.The English translations were added here for easy reading.

Data	(1) DRCD	(2) ODSQA	(3) DRCD-TTS	(4) DRCD-backtrans
D	...廣州屬亞熱帶季風海洋性氣候，氣候濕熱易上火的環境使飲涼茶成爲廣州人常年的一個生活習慣。 "Guangzhou has a subtropical monsoon maritime climate. Drinking cool tea has become a daily habit of Guangzhou people for a long time due to the humid and hot environment."	...廣州屬亞熱帶季風海洋性氣候，氣候濕熱，易上火的環境**時，應嚴查**成爲廣州人常年的一個生活習慣。 "Guangzhou has a subtropical monsoon maritime climate. **When** the environment is hot and humid, **necessarily strictly examination** has become a daily habit for Guangzhou people for a lone time."	...廣州屬亞熱帶季風海洋性氣候，氣候，**是誠意上火的環境適應量**，茶成廣州人常年的一個生活習慣。 "Guangzhou has a subtropical monsoon maritime climate. Climate, **being a sincere and hot humid adaptation to the environment**, tea has become a daily habit for Guangzhou people for a lone time."	...廣州屬亞熱帶季風海洋性氣候，氣候炎熱潮濕，是廣州人喝茶的共同習慣。 "Guangzhou has a subtropical monsoon maritime climate. The climate is hot and humid, which is a common habit of Guangzhou people when drinking tea."

DRCD is around 160,000 and the character vocabulary size is around 6,200. We experimented with both words and characters.

- **Implementation Details**

  Chinese word embeddings: We pre-train a Fasttext[36] model on words of traditional Chinese Wikipedia articles[5] segmented by jieba-zh-TW. This model can handle Out-of-Vocabulary words with character n-grams. The word embeddings in all our experiments were initialized from this 300 dimensional pre-trained fasttext model and fixed during training for both translated English word and Chinese word. This model is crucial to the performance of the qeustion answering models according to our experimental results.

  Chinese character embeddings: We pre-train a skip-gram model on characters of traditional Chinese Wikipedia articles using Gensim[6].

5.2. Baselines

We chose several competitive reading comprehension models here. The models are listed as follow:

- **BiDirectional Attention Flow (BiDAF)** [37]: In BIDAF, both character-level and word-level embeddings are incorporated. A Bi-directional attention flow mechanism, which computes attentions in two directions: from context to query as well as from query to context is introduced to obtain a query-aware context representation.
- **R-NET** [38]: In R-NET, the dependency in long context is captured more than plain recurrent neural network. A self-matching mechanism is introduced to dynamically refine context representation with information from the whole context.
- **QANet** [13]: There are completely no recurrent networks in QANet. Its encoder is composed of exlusively of convolution and self-attention. The intuition is that convolution components model local interactions and self-attention components model global interactions. Due to the removal of recurrent networks, it's training speed is 5x faster than BiDAF when reaching the same performance on SQuAD dataset.
- **FusionNet** [39]: There are mainly two contributions in FusionNet. First is the **History of Word**, in which all representations of a word from lowest level word embedding to the highest semantic level are concatenated to be the final representation of this word. Second is the **Fully-aware Multi-level Attention Mechanism**, which captures the complete information in one text (such as a question) and exploits it in its counterpart (such as a context or passage) layer by layer.
- **Dr.QA** [40]: Dr.QA is a rather simple neural network architecture compared to the previous introduced models. It basically is composed of multi-layer bidirectional long short-term memory networks. It utilizes some linguistic features such as part-of-speech tagging and name entity recognition.

In our task, during testing stage, all the baseline QA models take into a machine-transcribed spoken document and a machine-transcribed spoken question as input, and the output is an extracted span from the ASR transcription of docu-

[5]Wikipedia articles: https://dumps.wikimedia.org/zhwiki/

[6]Genism: https://radimrehurek.com/gensim/models/word2vec.html

Table 2. Data statistics of ODSQA, DRCD-TTS and DRCD-backtrans. The average document length and the average question length are denoted as Avg D Len and Avg Q Len respectively and they stand for the total numbers of Chinese characters. Since training with noisy data and different speakers will make QA model more robust during testing, the number of speakers is large. ODSQA testing set is denoted as ODSQA-test.

Subsets	QApair	Hours	M-spkrs	F-Spkrs	WER(%)	WER-Q(%)	Avg D Len	Avg Q Len
(1) ODSQA-test	1,465	25.28	7	13	19.11	18.57	428.32	22.08
(2) DRCD-TTS	16,746	–	–	–	33.63	–	332.80	20.53
(3) DRCD-backtrans	15,238	–	–	–	45.64	–	439.55	20.75

ment. We train these baseline QA models on DRCD training set and compare the performance between DRCD dev set and ODSQA testing set.

5.3. Artificially Generated Corpus

It is reported that training on transcriptions with ASR errors are better than training on text.[11], so we conduct the following procedures to generate transcriptions of spoken version DRCD. First, we used iFLYTEK Text-to-Speech system [7] to generate the spoken version of the articles in DRCD. Then we utilized iFLYTEK ASR system to obtain the corresponding ASR transcriptions. In this corpus, we left the questions in the text form. If the answer to a question did not exist in the ASR transcriptions of the associated article, we removed the question-answer pair from the corpus. This artificially generated corpus is called **DRCD-TTS** and its data statistics is shown in row(2) of Table 2.

5.4. Back-translation Corpus

To improve the robustness to speech recognition errors of QA model, we augmented DRCD training dataset with back-translation. We conduct the following procedures to generate an augmented training set. First, the DRCD training set is translated using Google Translation system into English. Then this set is translated back into Chinese through Google Translation system. We chose English as the pivot language here because English is the most common language and the translation quality is probably the best. Because the task is extraction-based QA, the ground-truth answer must exist in the document. Therefore, we removed the examples which cannot fulfill the requirement of extraction-based QA after translation. This resulting dataset is called **DRCD-backtrans** and its statistics is shown in row(3) of Table 2.

5.5. Result

First of all, we show the performance of the baseline models that are briefly introduced in subsection 5.2. All the QA models were trained on DRCD then test on DRCD dev set and ODSQA testing set respectively to compare the performance between text documents and spoken documents.

Secondly, we compare the performance of QA models with and without the proposed pingyin sequence embedding. Thirdly, we show how co-training with **DRCD-TTS** and **DRCD-backtrans** benefit. Last but not the least, we compare the performance between spoken question and text question.

Investigating the Impact of ASR Errors. We trained five reading comprehension models mentioned in Section 5.2 on the DRCD training set and these five models are tested on DRCD dev set and ODSQA testing set. In the following experiments, we do not consider the spoken documents whose answers do not exist in the ASR transcriptions because the model can never obtain the correct answers in these cases. To make the comparison fair, the DRCD dev set are filtered to contain only the same set of examples. As shown in Table 3, across the five models with char-based input, the average F1 score on the text document is 81.05%. The average F1 score fell to 63.67% when there are ASR errors. Similar phenomenon is observed on EM. The impact of ASR errors is significant for machine comprehension models. Since the author of BiDAF released its source code[8] and its decent performance over ODSQA testing set, we use it as the base model with char-based input for further experiments.

Mitigating ASR errors by Subword Units. We utilized an open-sourced chinese mandarin lexicon tool [9] to convert each word into sequence of pingyin-tokens and then fed the ping-yin tokens into Pingyin-CNN network to obtain pingyin-token sequence embedding. In this work, we didn't utilize tone information in pingyin-tokens. We leave it as a future work. The network details are listed as follow: pingyin-token embedding size 6, filter size 3x6 and numbers of filters 100. Different from [41] using one-hot vector, we choose distributed representation vectors to represent subword units. The experimental results with and without the proposed sub-word unit approach are in Table 4. We see from Table 4, using the combination of word embedding and the proposed pingyin sequence embedding is better than just word embedding (row (b)(d)(f)(h)(j)(l) vs. (a)(c)(e)(g)(i)(k)). The average EM score is improved by 1.3 by using pingyin

[7] iFLYTEK Text-to-Speech system
https://www.xfyun.cn/doccenter/tts

[8] BiDAF:Bi-directional Attention Flow for Machine Comprehension
https://github.com/allenai/bi-att-flow

[9] DaCiDian : an open-sourced chinese mandarin lexicon for automatic speech recognition(ASR)
https://github.com/aishell-foundation/DaCiDian

Table 3. Experiment results for state-of-the-art QA models demonstrating degrading performance under spoken data. All models were trained on the full DRCD training set. FusionNet is denoted by F-NET. DRCD dev set and ODSQA testing set are denoted by DRCD-dev and ODSQA-test, respectively.

MODEL	DRCD-dev		ODSQA-test	
	EM	F1	EM	F1
BiDAF-word(a)	56.45	70.57	39.38	55.1
BiDAF-char(b)	70.23	81.65	55.29	67.16
R-NET-word	70.38	79.25	36.68	46.55
R-NET-char	69.90	79.49	43.44	55.83
QAnet-word	69.83	78.33	49.80	59.35
QAnet-char	70.78	80.83	46.52	59.11
Dr.QA-word	63.21	74.11	41.39	54.28
Dr.QA-char	70.24	81.19	56.22	68.99
F-Net-word	57.54	70.86	45.39	57.40
F-Net-char	71.33	82.12	47.98	67.26
Average-word	63.48	74.62	42.52	54.53
Average-char	70.49	81.05	49.89	63.67

sequence embedding over ODSQA testing set.

Data augmentation. To improve the robustness to speech recognition errors of QA models, we augmented training data DRCD with DRCD-TTS and DRCD-backtrans. The experiment results are shown in Table 4. We can see from Table 4, training with the combination of DRCD and DRCD-backtrans or training with the combination of DRCD and DRCD-TTS are all better than training with only DRCD (row (g)(i) vs. (a) and row(h)(j) vs. row(b)). And finally training with the combination of DRCD, DRCD-TTS and DRCD-backtrans with pingyin sequence embedding obtains the best results (row(l)) which is better than baseline (row (a)) by almost 4 F1 score. Therefore, data augmentation proves to be helpful in boosting performance.

Comparison Between Text Question and ASR Transcribed Question. ASR errors on question will affect the reasoning of a QA model. In this part, we compare the performance between input with text questions and input with ASR-transcribed questions. We can see from Table 5, the average F1 score fell from 71.61% to 66.73% when there are ASR errors in question. Similar phenomenon is observed on EM. Once again, we can see that using pingyin sequence embedding brings improvement (row(h) vs (g)) even with text question as input.

6. CONCLUDING REMARKS

We release an SQA dataset, ODSQA. By testing several models, we found that ASR errors have catastrophic impact on SQA. We found that subword units bring consistent improvements over all the models. We also found that using back-translation and TTS to augment the text-based QA training examples can help SQA. In the future work, we are collecting more data for SQA.

Table 4. Comparison experiments demonstrating that the proposed sub-word units improved EM/F1 scores over both DRCD-dev and ODSQA-test. We use BiDAF as our base model in all experiments. Furthermore, augmenting DRCD with DRCD-TTS and DRCD-backtrans also gain improvements. Training with the combination of DRCD and DRCD-backtrans, the combination of DRCD and DRCD-TTS and the combination of DRCD, DRCD-TTS and DRCD-bakctrans are denoted as DRCD+back, DRCD+TTS and DRCD+TTS+back respectively.

MODEL	DRCD-dev		ODSQA-test	
	EM	F1	EM	F1
DRCD (a)	70.23	81.65	55.29	67.16
+pingyin (b)	71.05	81.82	55.49	68.79
DRCD-TTS (c)	59.24	72.64	50.64	63.65
+pingyin (d)	61.36	74.22	51.74	64.59
DRCD-back (e)	58.56	72.31	46.55	61.52
+pingyin (f)	58.63	72.97	48.2	62.82
DRCD+TTS (i)	70.51	81.85	55.97	69.31
+pingyin (j)	71.53	82.42	56.65	69.45
DRCD+back (g)	71.39	82.28	55.29	68.49
+pingyin (h)	71.8	82.4	57.6	69.26
DRCD+TTS+back (k)	72.21	82.8	57.61	70.29
+pingyin (l)	**72.76**	**83.15**	**59.52**	**71.01**
Average (m)	67.02	78.92	53.55	66.73
Average-pingyin (n)	67.85	79.49	54.86	67.65

Table 5. Comparison experiments between input with text question and input with transcribed question. We use BiDAF as our base model in all experiments.

MODEL	Text-Q		Spoken-Q	
	EM	F1	EM	F1
DRCD (a)	59.63	72.02	55.29	67.16
+pingyin (b)	61.47	72.93	55.49	68.79
DRCD-TTS (c)	54.43	67.18	50.64	63.65
+pingyin (d)	55.39	68.12	51.74	64.59
DRCD-back (a)	52.45	67.13	46.55	61.52
+pingyin (b)	53.41	68.57	48.2	62.82
DRCD+TTS (i)	61.95	73.78	55.97	69.31
+pingyin (j)	62.43	74.3	56.65	69.45
DRCD+back (c)	62.22	74.33	55.29	68.49
+pingyin (d)	62.7	74.81	57.6	69.26
DRCD+TTS+back (e)	63.11	75.27	58.29	69.94
+pingyin (f)	64.54	75.63	59.52	70.95
Average (g)	58.96	71.61	53.55	66.73
Average-pingyin (h)	59.99	72.39	54.86	67.65

7. REFERENCES

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

[2] Pere R Comas i Umbert, Jordi Turmo Borràs, and Lluís Màrquez Villodre, "Spoken question answering," .

[3] Pere R Comas Umbert, "Factoid question answering for spoken documents," 2012.

[4] Jordi Turmo, Pere R Comas, Sophie Rosset, Lori Lamel, Nicolas Moreau, and Djamel Mostefa, "Overview of qast 2008," in *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 2008, pp. 314–324.

[5] Pere R Comas, Jordi Turmo, and Lluís Màrquez, "Sibyl, a factoid question-answering system for spoken documents," *ACM Transactions on Information Systems (TOIS)*, vol. 30, no. 3, pp. 19, 2012.

[6] 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.

[7] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi, "Learning knowledge graphs for question answering through conversational dialog," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 851–861.

[8] 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.

[9] Wei Fang, Juei-Yang Hsu, Hung-yi Lee, and Lin-Shan Lee, "Hierarchical attention model for improved machine comprehension of spoken content," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 232–238.

[10] Yu-An Chung, Hung-Yi Lee, and James Glass, "Supervised and unsupervised transfer learning for question answering," *arXiv preprint arXiv:1711.05345*, 2017.

[11] 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.

[12] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," 2004.

[13] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[14] Matthew Richardson, Christopher JC Burges, and Erin Renshaw, "Mctest: A challenge dataset for the open-domain machine comprehension of text," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 193–203.

[15] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy, "Race: Large-scale reading comprehension dataset from examinations," *arXiv preprint arXiv:1704.04683*, 2017.

[16] 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.

[17] C Lawrence Zitnick, Ramakrishna Vedantam, and Devi Parikh, "Adopting abstract images for semantic scene understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 627–638, 2016.

[18] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[20] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler, "What are you talking about? text-to-image coreference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3558–3565.

[21] David L Chen and William B Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 190–200.

[22] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2634–2641.

[23] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[24] 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.

[25] Kenney Ng and Victor W Zue, "Subword unit representations for spoken document retrieval," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[26] Charl van Heerden, Damianos Karakos, Karthik Narasimhan, Marelie Davel, and Richard Schwartz, "Constructing sub-word units for spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5780–5784.

[27] Marijn Huijbregts, Mitchell McLaren, and David Van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4436–4439.

[28] Weidong Qu and Katsuhiko Shirai, "Using machine learning method and subword unit representations for spoken document categorization," in *Sixth International Conference on Spoken Language Processing*, 2000.

[29] Carolina Parada, Mark Dredze, Abhinav Sethy, and Ariya Rastrow, "Learning sub-word units for open vocabulary speech recognition," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 712–721.

[30] Xiaodong He, Li Deng, Dilek Hakkani-Tur, and Gokhan Tur, "Multi-style adaptive training for robust cross-lingual spoken language understanding," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8342–8346.

[31] Shyam Upadhyay, Manaal Faruqui, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck, "(almost) zero-shot cross-lingual spoken language understanding," 2018.

[32] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata, "Learning to paraphrase for question answering," *arXiv preprint arXiv:1708.06022*, 2017.

[33] 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.

[34] Xiang Zhang, Junbo Zhao, and Yann LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[35] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush, "Character-aware neural language models.," in *AAAI*, 2016, pp. 2741–2749.

[36] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[37] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[38] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou, "Gated self-matching networks for reading comprehension and question answering," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, vol. 1, pp. 189–198.

[39] Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen, "Fusionnet: Fusing via fully-aware attention with application to machine comprehension," *arXiv preprint arXiv:1711.07341*, 2017.

[40] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes, "Reading wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.

[41] Xu Li, Zhiyong Wu, Helen M Meng, Jia Jia, Xiaoyan Lou, and Lianhong Cai, "Phoneme embedding and its application to speech driven talking avatar synthesis.," in *INTERSPEECH*, 2016, pp. 1472–1476.