SPOKEN MULTIPLE-CHOICE QUESTION ANSWERING USING MULTIMODAL CONVOLUTIONAL NEURAL NETWORKS

Shang-Bao Luo[1]*, Hung-Shin Lee*[2]*, Kuan-Yu Chen*[1] *and Hsin-Min Wang*[2]

[1]National Taiwan University of Science and Technology, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

In a spoken multiple-choice question answering (MCQA) task, where passages, questions, and choices are given in the form of speech, usually only the auto-transcribed text is considered in system development. The acoustic-level information may contain useful cues for answer prediction. However, to the best of our knowledge, only a few studies focus on using the acoustic-level information or fusing the acoustic-level information with the text-level information for a spoken MCQA task. Therefore, this paper presents a hierarchical multistage multimodal (HMM) framework based on convolutional neural networks (CNNs) to integrate text- and acoustic-level statistics into neural modeling for spoken MCQA. Specifically, the acoustic-level statistics are expected to offset text inaccuracies caused by automatic speech recognition (ASR) systems or representation inadequacy lurking in word embedding generators, thereby making the spoken MCQA system robust. In the proposed HMM framework, two modalities are first manipulated to separately derive the acoustic- and text-level representations for the passage, question, and choices. Next, these clever features are jointly involved in inferring the relationships among the passage, question, and choices. Then, a final representation is derived for each choice, which encodes the relationship of the choice to the passage and question. Finally, the most likely answer is determined based on the individual final representations of all choices. Evaluated on the data of "Formosa Grand Challenge – Talk to AI", a Mandarin Chinese spoken MCQA contest held in 2018, the proposed HMM framework achieves remarkable improvements in accuracy over the text-only baseline.

***Index Terms*—** Spoken multiple-choice question answering, acoustic information, hierarchical multistage multimodal

1. INTRODUCTION

Machine reading comprehension (MRC) is an important task in the domain of natural language processing. The goal is to let the machine read the text and answer relevant questions based on the understanding of the text, like humans. Classic MRC tasks are performed on the plain text and can be classified into three main types, namely *answer span* [1, 2], i.e., finding the location of the answer in the paragraph, *cloze-style* [3, 4], i.e., inferring the missing entity (answer) from all possible entities appearing in the passage, and *multiple-choice* [5, 6], i.e., selecting the most likely answer from the multiple choices. Celebrated methods include BERT [7] and QANet [8] for locating the answer spans, attention-over-attention neural networks (AoA) [9] and ReasoNet [10] for filling in blanks, as well as various commonly used word embedding methods [11–14], to name a few.

In addition to text-based MRC, we have seen many personal voice assistant applications on a variety of mobile and home devices. This can be considered a special case of spoken question answering (SQA), i.e., questions are in the form of speech while the content can be in text or speech [15–18]. On the other hand, the development of multimedia technology and the popularity of video/audio sharing websites and social networks have also led to significant growth in spoken content nowadays. This has also increased the demand for machine comprehension of spoken content. Compared with text question answering, SQA needs to deal with ASR errors, which inevitably affect the performance of SQA systems. However, the acoustic cues embedded in the speech also provide additional semantic information that is not covered by the transcribed text. Therefore, if we can find some ways to make good use of acoustic information, we may improve the performance of SQA systems.

In this paper, we focus on the spoken multiple-choice question answering (MCQA) task, where passages, questions, and choices are all given in speech. A hierarchical multistage multimodal (HMM) framework based on convolutional neural networks (CNNs) is proposed to jointly utilize the text and acoustic characteristics. The two modalities are first manipulated to separately capture the representative acoustic-level and text-level features from the given passage, question, and each choice, respectively. Then, the resulting representations are simultaneously involved in training a multistage neural MCQA model, from which a final reading vector is derived for each tuple of passage, question, and choice. Accordingly, the most likely answer can be inferred from these final reading vectors associated with individual choices. In this paper, two kinds of acoustic features, namely filter banks (FBANKs)

and Mel-frequency cepstral coefficients (MFCCs), are explored. Evaluated on the data of "Formosa Grand Challenge – Talk to AI", a Mandarin Chinese spoken MCQA contest held in 2018, our system can achieve, when paired with the MFCC and FBANK features, relative accuracy improvement of 4.36% and 4.09%, respectively, compared to the text-only baseline system.

2. RELATED WORK

In a text-based MCQA task [19–23], the input to the model includes a passage, a question, and several answer choices. The passage usually consists of several sentences, while the question and each answer choice are a single sentence. The question answering model is designed to select a correct answer from multiple choices based on the information given in the passage and question. Along the line of research, previous works mainly focused on exploring textual cues, such as lexical and/or syntactic information embedded in the passage, question, and choices, to infer the answer to a given question [24–28]. Thanks to the rapid development of deep learning techniques, many studies have turned to the construction of neural MCQA models. Classic methods include the hierarchical attention-based CNN [29], the parallel-hierarchical neural model [30], and the hierarchical attention flow model [31], to name just a few. Although several elaborative mechanisms have been proposed based on deep neural networks, the query-based attention CNN (QACNN) model [19] can be considered as a representative. Specifically, QACNN is stacked with a similarity mapping layer, a QACNN layer, and a prediction layer. The similarity mapping layer is composed of an embedding layer and a comparison layer. The embedding layer is used to represent the passage, question, and choice with word vectors, and the comparison layer is to get the location relationship from the passage-question similarity matrix and the passage-choice similarity matrix. The QACNN layer, which is used to learn the location relationship pattern, contains a two-stage CNN with a question-based attention mechanism. In the first stage, each passage sentence is given a word-level feature first and then represented by a question-based sentence feature and a choice-based sentence feature. In the second stage, for each choice, the choice-based sentence features are pooled to a final choice-answer feature according to the sentence-level attention map computed from the question-based sentence features. The prediction layer is a fully-connected feed-forward network to compute the probability for each choice based on its choice-answer feature and output the most likely choice.

3. THE HIERARCHICAL MULTISTAGE MULTIMODAL (HMM) FRAMEWORK

Opposite to the text-based MCQA task, in the spoken MCQA task, the passage, question, and choices are all in speech. A naive but easy to implement the solution is to first transcribe these speech utterances into text using an ASR system. Thereafter, a classic method (e.g., QACNN) can be readily applied to the auto-transcribed text. This strategy only considers the auto-transcribed text, but it is obvious that the acoustic-level information may contain useful cues for answer prediction. Accordingly, in this paper, we focus on a systematic approach that combines acoustic and text-level cues with a CNN-based hierarchical multistage mutimodal (HMM) framework to improve the performance of the spoken MCQA task.

As shown in Fig. 1, the proposed HMM architecture consists of four stages, namely an acoustic-level similarity layer, a text-level similarity layer, an attention-based CNN layer, and a prediction layer.

In the **acoustic-level similarity layer**, we first quantify the passage, question, and *s*-th choice into a set of acoustic features $AP \in \mathbb{R}^{d\times L}$, $AQ \in \mathbb{R}^{d\times M}$, and $AC_s \in \mathbb{R}^{d\times N}$, where d is the feature dimension, and L, M, and N are the lengths of the passage, question, and *s*-th choice. Then, a deep CNN is used to extract the temporal information in the audio streams (i.e., AP, AQ, and AC_s), and maximum pooling ($MaxPool$) is applied to obtain the most salient information:

$$E_P^A = MaxPool(W_3^{CNN} \otimes (W_2^{CNN} \otimes (W_1^{CNN} \otimes AP))) \quad (1)$$

$$E_Q^A = MaxPool(W_3^{CNN} \otimes (W_2^{CNN} \otimes (W_1^{CNN} \otimes AQ))) \quad (2)$$

$$E_{C_s}^A = MaxPool(W_3^{CNN} \otimes (W_2^{CNN} \otimes (W_1^{CNN} \otimes AC_s))) \quad (3)$$

where $\otimes$ denotes the operation of convolution, W_1^{CNN}, W_2^{CNN}, and W_3^{CNN} are the parameters of the convolution layers, $E_P^A \in \mathbb{R}^{f\times I}$, $E_Q^A \in \mathbb{R}^{f\times J}$ and $E_{C_s}^A \in \mathbb{R}^{f\times K}$ are the resulting acoustic-level embedding features for the passage, question, and *s*-th choice, respectively. These features can be treated as acoustic frame-based statistics. Note that the architecture configurations of the acoustic-level similarity layer are specially designed to make I, J, and K same as the numbers of words in the passage, question, and *s*-th choice, respectively. Next, we infer the relationship between the passage and the question (i.e., E_P^A and E_Q^A) as well as the relationship between the passage and the *s*-th choice (i.e., E_P^A and $E_{C_s}^A$). To do so, fully-connected feed-forward networks are used to translate these features into a common space, and then an attention mechanism [32] is used to create a passage-to-question matrix $PQ^A \in \mathbb{R}^{I\times J}$ and a passage-to-choice matrix $PC_s^A \in \mathbb{R}^{I\times K}$:

$$PQ^A = softmax((W_1^{FC} E_P^A)^T \times (W_2^{FC} E_Q^A)) \quad (4)$$

$$PC_s^A = softmax((W_1^{FC} E_P^A)^T \times (W_2^{FC} E_{C_s}^A)) \quad (5)$$

where W_1^{FC} and W_2^{FC} are the parameters of the fully-connected feed-forward networks, and the softmax function ($softmax$) is performed across the rows of these matrices. Consequently, the *i*-th row of the matrix PQ^A (or PC_s^A)

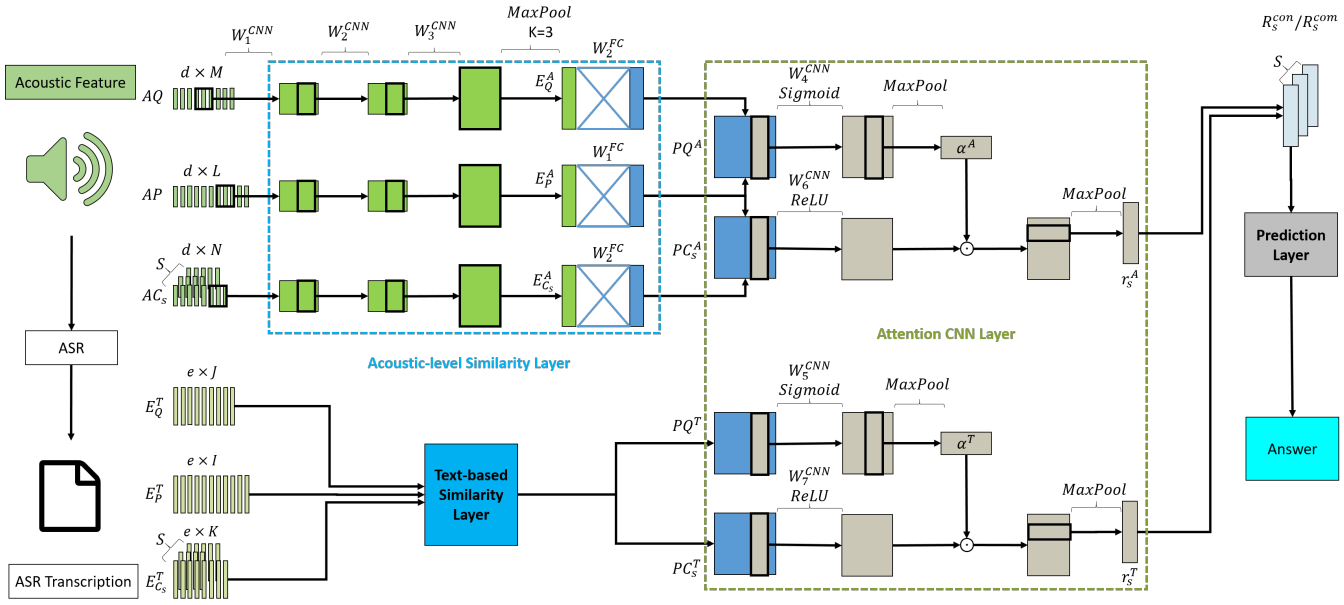


Fig. 1. The overview of the proposed hierarchical multistage multimodal (HMM) framework.

signifies which question spans (or spans in the s-th choice) are most relevant to the i-th passage span.

The **text-based similarity layer** is applied to the auto-transcribed text for the passage, question, and s-th choice represented as matrices $E_P^T \in \mathbb{R}^{e\times I}$, $E_Q^T \in \mathbb{R}^{e\times J}$, and $E_{C_s}^T \in \mathbb{R}^{e\times K}$, where e is the dimension of the pre-trained fix-dimensional word embedding vector, and I, J, and K are the numbers of words in the passage, question, and s-th choice, respectively. Based on these word-level semantic features, the text-based similarity layer can measure the similarity between the passage and the question as well as the similarity between the passage and the s-th choice. A straightforward strategy is to create a passage-to-question matrix $PQ^T \in \mathbb{R}^{I\times J}$ and a passage-to-choice matrix $PC_s^T \in \mathbb{R}^{I\times K}$ by computing the cosine similarity between each pair of words:

$$PQ^T = cos(E_P^T, E_Q^T) \tag{6}$$

$$PC_s^T = cos(E_P^T, E_{C_s}^T). \tag{7}$$

The attention CNN layer is responsible for further fusing the information in the passage, question, and s-th choice and generating the choice-dependent reading vectors, in the word-level and acoustic-level granularities, respectively. The passage-to-question matrix (i.e., PQ^A or PQ^T) is processed by a convolution layer, a Sigmoid activation function, and a maximum pooling function to generate an essential vector:

$$\alpha^A = MaxPool(Sigmoid(W_4^{CNN} \otimes PQ^A)) \tag{8}$$

$$\alpha^T = MaxPool(Sigmoid(W_5^{CNN} \otimes PQ^T)) \tag{9}$$

where W_4^{CNN} and W_5^{CNN} are the parameters of the convolution layers. Here, 1-D convolution and maximum pooling are both performed along with the passage words. The passage-to-choice matrix (i.e., PC_s^A or PC_s^T) is processed by a convolution layer and a ReLU activation function, scaled by the essential vector (i.e., α^A or α^T), and transformed to a reading vector (i.e., r_s^A or r_s^T) by maximum pooling:

$$r_s^A = MaxPool(ReLU(W_6^{CNN} \otimes PC_s^A) \odot \alpha^A) \tag{10}$$

$$r_s^T = MaxPool(ReLU(W_7^{CNN} \otimes PC_s^T) \odot \alpha^T) \tag{11}$$

where $\odot$ denotes element-wise multiplication, and W_6^{CNN} and W_7^{CNN} are the parameters of the CNN layers. Here, maximum pooling is performed along with the choice words. r_s^A and r_s^T are called "reading vectors" because they convey

Table 1. Configurations of the proposed HMM framework.

Convolutional Neural Network	Filter Number	Kernel Size
W_1^{CNN}	128	3
W_2^{CNN}	128	3
W_3^{CNN}	256	3
W_4^{CNN}	256	c.f. Table 3
W_5^{CNN}	256	c.f. Table 3
W_6^{CNN}	256	c.f. Table 3
W_7^{CNN}	256	c.f. Table 3
Fully-connected Feed-forward	**Size of Output Layer**	
W_1^{FC}	256	
W_2^{FC}	256	
W_3^{FC}	256	
W_4^{FC}	1	

the reciprocal information among the passage, question, and *s*-th choice.

In the last stage, **the prediction layer** selects the most relevant choice for the given passage and question by considering multimodal features. A naive strategy is to concatenate the derived acoustic- and text-based reading vectors together:

$$R_s^{con} = [r_s^A; r_s^T]. \tag{12}$$

Alternatively, we can linearly combine the two vectors:

$$R_s^{com} = \beta \times r_s^A + (1 - \beta) \times r_s^T \tag{13}$$

where β is a weighting factor. β was set to 0.9 in this study. On top of the fusion vector (i.e., R_s^{con} and R_s^{com}), the relevance score for the *s*-th choice is determined by a two-layer fully-connected feed-forward network:

$$P(R_s) = \frac{exp(W_4^{FC}(tanh(W_3^{FC}R_s)))}{\sum_{s'=1}^{S} exp(W_4^{FC}(tanh(W_3^{FC}R_{s'})))} \tag{14}$$

where $R_s \in \{R_s^{con}, R_s^{com}\}$, S denotes the number of choices, W_3^{FC} and W_4^{FC} are the parameters of the fully-connected feed-forward layers. Finally, the choice with the largest relevance score will be selected as the answer. Note that the nonlinear normalization function in Eq. (14) is only used for training; for testing, there is no need to normalize the scores.

4. EXPERIMENTS

4.1. Experimental Setup

We used the 2018 Formosa Grand Challenge (FGC) dataset[1] in the experiments. The challenge is a Mandarin Chinese spoken MCQA task. Each passage-question-choices (PQC) set contains a passage, a question, and 4 choices, among which only one choice is the correct answer. The domain of the FGC dataset is very diverse, including science, news, and literature, to name a few. There are 9,083 PQC sets, 1,430 sets for development, 977 sets for testing, and 6,676 sets for training. The average number of frames for each passage, question, and choices audio file are 22,538, 2,230, and 1,476.

ASR. Our ASR system was built up using the Kaldi toolkit [33], where the acoustic model was trained based on TDNN-F with lattice-free MMI [34, 35], followed by model refinement with sMBR [36], with 487 hours of TV and radio broadcasting speech. In audio processing, spectral analysis was applied to a 25 ms frame of speech waveform every 10 ms. For each acoustic frame, 40 MFCCs derived from 40 FBANKs, plus 3 pitch features, were used for ASR and for our proposed HMM framework. Utterance-based mean subtraction was applied to these features. The lexicon contained 91,573 Chinese words. The word-based trigram language model was trained with Kneser-Ney backoff smoothing using the SRILM toolkit [37]. The recurrent neural network language model (RNNLM) was used for lattice rescoring [38]. The training corpus was compiled from PTT [2] articles (2018) and CNA news stories (2006 ~ 2010) [39]. The character error rate (CER) of our ASR system is about 7.79%.

Implementation Details. We used fasttext modeling [13] to train word embeddings on the same corpus for language model training. The dimension of the word embedding (i.e., e) was set to 300. The numbers of words for a passage, question, and choice were limited to 1,000, 50, and 50, respectively. The numbers of acoustic frames for a passage, question, and choice were limited to 24,000, 1,200, and 1,200, respectively. The framework was implemented by PyTorch [40], and the model parameters were optimized by the Adam method [41]. We used Chinese characters instead of words as the basic semantic unit to avoid the word segmentation error issue. The detailed configurations of the proposed HMM framework are summarized in Table 1.

Table 2. Performance (in accuracy (%)) of different systems.

	Dev	**Test**
Choice Length		
Longest	37.97	30.30
Shortest	19.09	21.70
Choice Similarity		
Passage-choice	26.01	25.59
Question-choice	46.71	41.56
QACNN(ASR)	62.94	72.16
QACNN(Oracle)	69.03	80.82
HMM		
w/ Weight Sharing+R^{con}+MFCCs	63.43	73.90
w/ Weight Sharing+R^{con}+FBANKs	63.50	71.55
w/ Weight Sharing+R^{com}+MFCCs	63.57	72.67
w/ Weight Sharing+R^{com}+FBANKs	64.06	73.08
w/o Weight sharing+R^{con}+MFCCs	64.06	72.88
w/o Weight sharing+R^{con}+FBANKs	63.99	72.57
w/o Weight sharing+R^{com}+MFCCs	64.41	73.08
w/o Weight sharing+R^{com}+FBANKs	63.78	74.72

4.2. Results and Discussions

In the first set of experiments, we compared the proposed methods with several baselines. The experimental results are summarized in Table 2. The most naive baseline is to choose the longest choice or the shortest choice as the answer (denoted by "Choice Length" in Table 2). This method could be even worse than a random guess. A simple way of leveraging word embeddings is to represent a passage/question/choice

[1] Formosa Grand Challenge - Talk to AI: https://fgc.stpi.narl.org.tw/activity/techai2018

[2] PTT : https://www.ptt.cc/index.html

Table 3. Performance (in accuracy (%)) of the HMM framework with different configurations in the attention CNN layer.

	w/ Weight Sharing [MFCCs / FBANKs]		w/o Weight Sharing [MFCCs / FBANKs]	
Concatenation Method (R^{con})	**Dev**	**Test**	**Dev**	**Test**
[1,2,7]	63.15 / 63.15	73.18 / 72.26	63.22 / 63.71	74.10 / 73.29
[1,3,4]	63.01 / **63.50**	72.47 / **71.55**	62.73 / 63.22	73.59 / 72.88
[1,3,5]	**63.43** / 63.01	**73.90** / 73.08	**64.06** / 63.08	**72.88** / 72.26
[1,3,7]	63.01 / 62.66	73.90 / 72.36	63.01 / **63.99**	73.18 / **72.57**
Linear Combination Method (R^{com})	**Dev**	**Test**	**Dev**	**Test**
[1,2,7]	**63.57** / 63.22	**72.67** / 72.67	**64.41 / 63.78**	**73.08 / 74.31**
[1,3,4]	63.43 / 63.29	73.80 / 72.88	63.36 / 63.01	72.36 / 72.98
[1,3,5]	63.08 / 63.29	73.18 / 72.88	63.57 / 63.43	73.18 / 73.69
[1,3,7]	62.24 / **64.06**	73.18 / **73.08**	64.06 / 63.50	74.00 / 74.72

Table 4. Performance (in accuracy(%)) of the HMM framework using only the acoustic-level features [MFCCs / FBANKs].

	Dev	Test
[1,2,7]	35.66 / 35.03	29.27 / 31.32
[1,3,4]	36.29 / 37.83	29.27 / 29.80
[1,3,5]	36.29 / 36.85	29.27 / 30.71
[1,3,7]	36.15 / 36.92	28.76 / 30.09

by averaging the embeddings of all the words in the passage/question/choice. Then, we can choose the choice with the largest cosine similarity with the passage or the question as to the answer. The results, as denoted by "Choice Similarity" in Table 2, indicate that the relationship between the question and the choice is more effective than the relationship between the passage and the choice. The third baseline is the QACNN model [19]. As expected, QACNN performed much better than the "Choice Length" and "Choice Similarity" methods. The results demonstrate the ability of the QACNN model and the potential of neural methods.

For the proposed HMM framework, two acoustic-level features (i.e., MFCCs and FBANKs), two fusion vectors (i.e., R^{con} and R^{com}), and two model configurations (whether the attention CNN layers for the acoustic and text streams are shared or not) were investigated. First, we find that the proposed HMM framework outperforms all baseline systems in almost all cases, which signals that it can indeed make use of both acoustic-level and text-level statistics for answer prediction. Second, from the perspective of acoustic information, the results show that the MFCC and FBANK features are neck and neck with each other. Third, the linear combination method for fusing the acoustic-level and text-level reading vectors (i.e., R^{com}) outperforms the concatenation method (i.e., R^{con}) in most cases. Fourth, the experimental results do not suggest sharing the attention CNN layers for the acoustic and text streams. A preferred way is to train separate models to extract useful information from the acoustic and text streams, respectively.

In the second set of experiments, we looked into the impact of the size of the kernel used in the attention CNN layer in the proposed HMM framework on the performance. Take "[1,2,7]" for example, it means that each CNN (i.e., W_4^{CNN}, W_5^{CNN}, W_6^{CNN}, and W_7^{CNN}) in the attention CNN layer has three different sizes of kernels, namely 1, 2, and 7, and for each size, there are 256 filters (c.f. Table 1). From the results in Table 3, it is difficult to come up with a rule for setting the sizes of kernels for the CNNs in the attention CNN layer. Systemically determining the kernel size for a CNN is still an open issue and needs further investigation.

Finally, we explored the performance of the acoustic-level features in the spoken MCQA task. We used the acoustic-level features alone in the proposed HMM framework. The resulting model can be considered as an acoustic-based QACNN. The results are presented in Table 4. Comparing the results in Tables 2 and 4, it is clear that the acoustic-based QACNN is worse than the text-based QACNN, but can achieve comparable results to the naive word-based baseline systems (i.e., the "Choice Length" and "Choice Similarity" methods). In summary, the proposed HMM framework is deemed a preferable vehicle for utilizing acoustic-level and text-level characteristics in the spoken MCQA task.

5. CONCLUSIONS

In this paper, we have presented a hierarchical multistage multimodal (HMM) framework, which jointly considers the acoustic-level and text-level statistics for the spoken MCQA task. The proposed HMM framework has been evaluated on the 2018 Formosa Grand Challenge (FGC) dataset. The experimental results demonstrate its remarkable superiority than other strong baselines compared in the paper, thereby indicating the potential of the new spoken MCQA framework. For future work, we will explore additional cues, such

as sub-word information, in the proposed framework for the MCQA task. We also plan to extend the proposed HMM framework to other NLP-related tasks, such as retrieval and summarization, and evaluate the framework on other spoken question answering datasets.

6. ACKNOWLEDGEMENTS

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 108-2636-E-011-005 (Young Scholar Fellowship Program), and by the Project J367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

7. REFERENCES

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proc. EMNLP*, 2016, pp. 2383–2392.

[2] 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.

[3] Danqi Chen, Jason Bolton, and Christopher D Manning, "A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task," in *Proc. ACL*, 2016, pp. 2358–2367.

[4] Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester, "Who did What: A Large-Scale Person-Centered Cloze Dataset," in *Proc. EMNLP*, 2016, pp. 2230–2235.

[5] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi, "Are You Smarter than A Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension," in *Proc. CVPR*, 2017, pp. 4999–5007.

[6] Johannes Welbl, Nelson F Liu, and Matt Gardner, "Crowdsourcing Multiple Choice Science Questions," in *Proc. Workshop on Noisy User-generated Text*, 2017, pp. 94–106.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] 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.

[9] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu, "Attention-over-Attention Neural Networks for Reading Comprehension," in *Proc. ACL*, 2017, pp. 593–602.

[10] Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen, "Reasonet: Learning to Stop Reading in Machine Comprehension," in *Proc. KDD*, 2017, pp. 1047–1055.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning, "Glove: Global Vectors for Word Representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching Word Vectors with Subword Information," 2017, vol. 5, pp. 135–146.

[14] Benjamin Heinzerling and Michael Strube, "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages," in *Proc. LREC*, 2018.

[15] Pere R Comas Umbert, J. T. Borràs, and L. M. Villodre, *Spoken Question Answering*, Ph.D. thesis, Universitat Politècnica de Catalunya, 2012.

[16] Pere R. Comas i Umbert, *Factoid Question Answering for Spoken Documents*, Ph.D. thesis, Universitat Politècnica de Catalunya, 2012.

[17] Chia-Hsuan Lee, 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," in *Proc. Interspeech*, 2018, pp. 3459–3463.

[18] Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee, "Odsqa: Open-domain Spoken Question Answering Dataset," in *Proc. SLT*, 2018, pp. 949–956.

[19] Tzu-Chien Liu, Yu-Hsueh Wu, and Hung-Yi Lee, "Query-based Attention CNN for Text Similarity Map," *arXiv preprint arXiv:1709.05036*, 2017.

[20] Akshay Chaturvedi, Onkar Pandit, and Utpal Garain, "CNN for Text-Based Multiple Choice Question Answering," in *Proc. ACL*, 2018, pp. 272–277.

[21] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III, "A Neural Network for Factoid Question Answering over Paragraphs," in *Proc. EMNLP*, 2014, pp. 633–644.

[22] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha, "Finding the Right Facts in the Crowd : Factoid Question Answering over Social Media Categories and Subject Descriptors," in *Proc. WWW*, 2008, pp. 467–476.

[23] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou, "LSTM-based Deep Learning Models for Non-factoid Answer Selection," *arXiv preprint arXiv:1511.04108*, 2015.

[24] Matthew Richardson, Christopher JC Burges, and Erin Renshaw, "MCTEST: A Challenge Dataset for the Open-domain Machine Comprehension of Text," in *Proc. EMNLP*, 2013, pp. 193–203.

[25] Mrinmaya Sachan, Kumar Dubey, Eric Xing, and Matthew Richardson, "Learning Answer-entailing Structures for Machine Comprehension," in *Proc. IJCNLP*, 2015, pp. 239–249.

[26] Karthik Narasimhan and Regina Barzilay, "Machine Comprehension with Discourse Relations," in *Proc. IJCNLP*, 2015, pp. 1253–1262.

[27] Ellery Smith, Nicola Greco, Matko Bosnjak, and Andreas Vlachos, "A Strong Lexical Matching Method for the Machine Comprehension Test," in *Proc. EMNLP*, 2015, pp. 1693–1698.

[28] Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester, "Machine Comprehension with Syntax, Frames, and Semantics," in *Proc. IJCNLP*, 2015, pp. 700–706.

[29] Wenpeng Yin, Sebastian Ebert, and Hinrich Schütze, "Attention-based Convolutional Neural Network for Machine Comprehension," in *Proc. the Workshop on Human-Computer Question Answering*, 2016, pp. 15–21.

[30] Adam Trischler, Zheng Ye, Xingdi Yuan, Jing He, and Philip Bachman, "A Parallel-Hierarchical Model for Machine Comprehension on Sparse Data," in *Proc. ACL*, 2016, pp. 432–441.

[31] Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu, "Hierarchical Attention Flow for Multiple-choice Reading Comprehension," in *Proc. AAAI*, 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is All You Need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[33] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Luk Burget, Ondej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky Silovsky, Georg Stemmer, and Karel Vesely Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011.

[34] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohamadi, and Sanjeev Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.

[35] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[36] Karel Vesely, Arnab Ghoshal, Luk Burget, and Daniel Povey, "Sequence-discriminative Training of Deep Neural Networks," in *Proc. Interspeech*, 2013, pp. 2345–2349.

[37] Andreas Stolcke, "SRILM: An Extensible Language Modeling Toolkit," in *Proc. Interspeech*, 2002.

[38] Hainan Xu, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur, "A Pruned RNNLM Lattice-Rescoring Algorithm for Automatic Speech Recognition," in *Proc. ICASSP*, 2018, pp. 5929–5933.

[39] David Graff and Ke Chen, "Chinese Gigaword (LDC2003T09)," 2003.

[40] Adam Paszke, Gregory Chanan, Zeming Lin, Sam Gross, Edward Yang, Luca Antiga, and Zachary Devito, "Automatic Differentiation in PyTorch," in *Proc. NIPS*, 2017.

[41] Diederik P Kingma and Jimmy Lei Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015.