FAST TASK-SPECIFIC ADAPTATION IN SPOKEN LANGUAGE ASSESSMENT WITH META-LEARNING

Binghuai Lin[⋆] , *Liyuan Wang* [⋆]

[⋆]Smart Platform Product Department, Tencent Technology Co., Ltd, China

ABSTRACT

Automatic spoken language assessment plays an important role in assessing English proficiency of non-native learners, which involves tasks ranging from restricted tasks such as Repeat Sentences to more open-ended tasks such as unconstrained spontaneous speech. Traditional methods typically focus on specific task types and rely on a significant amount of human-labelled data. In this paper, we propose a fast adaptation framework with meta-learning for various task types in spoken language assessment under low-resource settings. To better adapt to tasks with different grading criteria, we incorporate a memory network acting as an external memory for these criteria. Experimental results based on data from different spoken language tests demonstrate the superiority of the proposed method to the baselines in Pearson correlation coefficient and accuracy when adapted to various task types, especially in low-resource settings.

Index Terms— Spoken language assessment, fast adaptation, low-resource, memory network, various tasks

1. INTRODUCTION

Automatic spoken language assessment plays an important role in computer-assisted language learning (CALL), which can help English-as-second-language (ESL) learners assess their English skills automatically. In the past, it focuses mainly on restricted tasks such as Repeat Sentences and Read aloud. For more comprehensive evaluation of spoken language skills, it has been extended to unconstrained tasks such as Picture Description, Short-answer Questions, Topic, and Opinion tasks [1]. Some examples of these tasks are shown in Figure 1.

Much previous work has been focused on restricted spoken tasks such as pronunciation assessment. Features used for pronunciation assessment are usually extracted from the hidden Markov model (HMM) of an automatic speech recognizer. HMM likelihood, posterior probability, and pronunciation duration features were proposed for pronunciation assessment in [2]. A variation of the posterior probability ratio called the Goodness of Pronunciation (GOP) [3] was proposed for pronunciation evaluation [4].

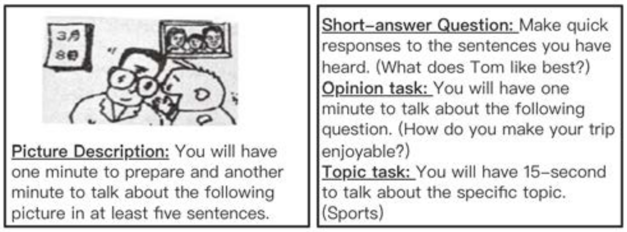
3月
8日
Picture Description: You will have one minute to prepare and another minute to talk about the following picture in at least five sentences.
Short-answer Question: Make quick responses to the sentences you have heard. (What does Tom like best?)
Opinion task: You will have one minute to talk about the following question. (How do you make your trip enjoyable?)
Topic task: You will have 15-second to talk about the specific topic. (Sports)

Fig. 1. Examples of different spoken language assessment tasks

With the development of automatic speech recognition (ASR) and spoken language processing, automated assessment of spoken language has been extended to more unconstrained tasks with more comprehensive features such as delivery and text features. Content features indicating the relevance and correctness of the facts contained in the response were considered for automatic scoring of spoken responses [5]. Speaking proficiency as well as grammatical and content accuracy was considered for automatic speech scoring [6]. Text-driven features (e.g., lowercased word n-grams) and speech-driven features (e.g., speech rate) were combined in automatic scoring of open-ended questions by support vector regressor (SVR) [7]. With the development of the neural network (NN), much work has been investigated to extract features by NN without complicated manual feature engineering. An attention-based model using Bidirectional Recurrent Neural Network (BiRNN) was designed to detect the off-topic spontaneous spoken response [8]. A siamese convolutional neural networks (CNN) combining both recognized transcripts and prompts, was proposed to score the spontaneous speech automatically [9]. NN approaches for automatic assessment of non-native spontaneous speech were proposed based on three attention-based Bi-Directional Long Short-Time Memory modules (BD-LSTM), which were used for extracting features relating to content, delivery, and language use features [10].

The aforementioned methods for spoken language assessment are designed for specific tasks, which are not appropriate or difficult for adaptation to other tasks. To design more adaptable scoring models, some work evaluated the reasonableness of spoken responses by considering the relationship

These authors annotated with ⋆ contributed equally to this work.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

between prompts or key points and the responses based on the network such as attention-based LSTM-RNN [11], which was proved to be a more generic model for the responses to both seen prompts and unseen prompts. However, these methods depend on significant amounts of labelled data and focus only on content-based features without considering other factors such as pronunciation, limiting their adaptability.

Human annotations for spoken language assessment tasks normally are difficult and expensive to collect. We can treat them as low-resource problems. Recently, many studies have been carried out to tackle this challenge in other areas. Low-resource ASR or neural machine translation was implemented by multi-lingual multi-task (MTL) pre-training and fine-tuning or a meta-learning framework [12, 13]. Inspired by these studies, we propose fast task-specific adaptation in spoken language assessment under a meta-learning framework called Model-Agnostic Meta-Learning (MAML) [14].

For fast adaptation to specific assessment tasks, we focus on two challenges : (1) adaptability to different tasks in spoken language assessment; and (2) data scarcity problem. We attempt to tackle the problem with a meta-learning framework. Considering part of the adaptability problem arising from various grading criteria, the memory network [15], an attention model over a large external memory, is utilized as an external learnable memory unit for these criteria. We compare the proposed method with previous traditional baselines as well as the MTL pre-training and fine-tuning-based method with different numbers of training data for different tasks. Experimental results have shown the effectiveness and superiority of the proposed method.

2. PROPOSED METHOD

We will introduce the proposed method from three aspects : (1) feature extraction module; (2) memory-network-based scoring module; (3) meta-learning-based training strategy.

2.1. Feature extraction module

The feature extraction module takes the audios (spoken responses), the prompts, which contain information that should be included in the answers (e.g., reading passages, listening passages, and sample responses) as input, and multimodal features containing text and delivery features as output.

The delivery features represent the quality of the pronunciation, fluency, and prosody of the spoken responses. First, we obtain the alignment information through a trained ASR model. For pronunciation assessment, we calculate the phoneme-level GOP based on the ASR acoustic model and the alignments [4]. We derive a phoneme representation vector by averaging GOP scores for each phoneme and obtain the sentence-level GOP by averaging the GOP of the phonemes in the sentence. For fluency and prosody assessment, we extract speech rate, pausing and timing information following previous work [7]. Language model (LM) scores and confidence scores derived from ASR are also included in the delivery features [16].

The text features are composed of syntax, content relatedness, and topic. Based on the trained ASR model, we derive the word hypotheses of the spoken responses, from which we can get syntactic features utilizing POS LMs [17]. For content relatedness, we evaluate how much the spoken response covers the sample responses as well as the semantic similarity between them based on pre-trained BERT [18]. We also calculate the topic vector of the spoken response based on latent Dirichlet allocation (LDA) [19].

2.2. Memory-network-based scoring module

The aforementioned features are normalized to zero means and one variances before fed into a scoring module as shown in Figure 2.

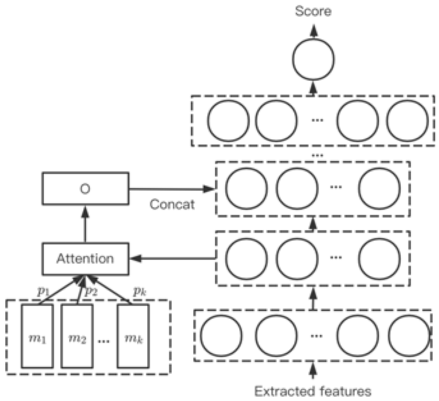

Fig. 2. Memory-network-based scoring network structure

The scoring module is composed of a DNN and an additional bank of embeddings. We refer them as criterion embeddings, which are designed to learn and store the grading criteria dynamically. We first map the input features through a linear transformation followed by a Tanh activation function to obtain the representation f with the dimensionality of d. We randomly initialize the criterion embedding matrix M with the dimensionality of $k \times d$. The attention weights between f (of size $1 \times d$) and criterion embeddings are defined in Eq. (1). Then, we sum the embeddings based on these weights as defined in Eq. (2). The weighted criterion embeddings are fused with the feature representations by concatenation. The concatenated representations are fed into multiple NN layers to calculate the final scores.

$$P = \mathrm{Softmax}(f^T M) \qquad (1)$$

$$O = \sum_{i=1}^{k} p_i m_i \qquad (2)$$

2.3. Meta-learning for scoring

Meta-learning is aimed to pre-train the models to get a good initialization for fast adaptation to new tasks with few iterations and less training data [14]. Given n tasks composed of various same or different task types $\mathrm{T} = \{t_1, t_2, ...t_n\}$, MAML obtains good initialization parameters θ^* of the scoring module capable of fast task-specific learning for the task t_d whether it is of new task type or not. The training loss L is the mean squared error (MSE) loss between the human and predicted scores. The algorithm is summarized in Algorithm 1. For ith iteration, we sample T_i tasks for meta-learning. For each task, we choose K samples for training (support samples) and L samples for development (query samples) to optimize the local parameters θ_i. Then, we sum over the validation loss of all tasks in the batch to update the global parameters θ. Once the global loss converges, we are ready to fine-tune the network for fast adaptation to the target task.

Algorithm 1: Scoring MAML

```
Input: $\mathrm{T} = t_1, t_2, ...t_n$: n tasks, $\alpha, \beta$: step size
       hyper-parameters, $t_d$: a specific task to adapt
       to
Output: learned parameters $\theta^d$
randomly initialize $\theta$
while not converged do
    Sample batch of tasks $T_i$
    for $t_j$ in $T_i$ do
        Compute local adapted parameters
        $\theta_i = \theta - \alpha \nabla L_{t_j}^{train}(\theta)$
    Update initialization parameter $\theta$
      $\theta = \theta - \beta \sum_{T_i} \nabla L_{t_j}^{valid}(\theta_i)$
Fine-tune $t_d$ with parameters $\theta^*$
```

3. EXPERIMENTS

3.1. Corpus

The speech scoring data consist of four different task types: Picture Description, Short-answer Questions, Topic task, and Opinion task, which come from one English oral test of the Chinese National Higher Education Entrance Examination. In this work, we split task types into seen and unseen types to test the adaptability to new tasks and new types. Specifically, we treat Picture Description, Short-answer Questions, and Topic task as seen types as they are in the training sets of meta-learning training, and Opinion task as an unseen type. Each seen task type contains four different tasks, ending up 12 tasks in total for meta-learning training. We test two different scenarios during fast adaptation stage, including a new task of seen type (Picture Description) and a new task of unseen type (Opinion task), to validate its adaptability. The statistics of the data are shown in Table 1.

The scoring ranges of Picture Description, Short-answer Questions, Topic task, and Opinion task are 0-3, 0-1, 0-5, and 0-5 (all with 0.5-grade interval), respectively, with the lowest score representing completely wrong answers and the highest score representing perfect responses. Three experts rated the responses, and the final grades were obtained by majority voting. The averaged inter-rater correlations calculated by Pearson correlation coefficient (PCC) between scores of one rater and average scores of the other two are 0.77. We normalize scores of different tasks to the range of 0 to 1 to fit them into the meta-learning framework.

Table 1. Statistics of data for meta-learning and fast adaptation testing

Stage	Task type	# of tasks	# of Training data (support samples)	# of Testing data (query samples)
Meta learning Training	Picture Description	4	50×4	150×4
	Short-answer Questions	4	50×4	150×4
	Topic task	4	50×4	150×4

Stage	Task type	# of tasks	# of Training data (fast adaptation)	# of Testing data
Fast adaptation Testing	Picture Description (new task of seen type)	1	50	1150
	Opinion task (new task of unseen type)	1	50	1150

3.2. Experimental setup

Based on the feature extraction module mentioned in section 2.1, we obtain delivery features with the dimensionality of 69, with 40 for phoneme GOP and sentence GOP, 27 for fluency and rhythm, and 2 for LM and confidence scores. The content-based features have the size of 52, consisting of 30 for syntactic features, 2 for content relatedness, and 20 for topic features. Thus, the total feature dimensionality is 121. The scoring module is composed of three layers. The first layer is composed of a fully connected (FC) layer with a size of 121×50 followed by a Tanh activation function. The memory network is composed of 5 distinctive criterion embeddings with the dimensionality of 1×50 for each. The second and third FC layers have the size of 100×50 and 50×1, respectively, with an additional Tanh activation function after the second one. The hyper-parameters α and β are 0.01 and 0.001, respectively.

The performance of the scoring model is evaluated by calculating PCC between predicted scores and expert labels as well as the accuracy, which is defined as ratios of low prediction errors being inside 0.5 (i.e., less than or equal to half a grade out) or inside 1.0 (i.e., less than or equal to a full grade out) of the expert grades.

3.3. Comparative study

First, we compare results with traditional full-training methods trained only with target task data, and MTL pre-training and task fine-tuning method (MTL-finetune) for the new task of seen type (Picture description). The architecture of the MTL-finetune method is the same as our proposed network shown in Figure 2. The tasks use the same input features as the proposed approach and share the whole network except for the last fully connected layer calculating the final scores. Then, we evaluate the method for the new task of unseen type, i.e. not appearing in the MAML-training (Opinion task). Finally, we explore the performance of the new task type (Opinion task) under different adaptation data sizes.

Two baselines of traditional full-training methods are compared here. The first one is based on the previous work [7], which feds the text and delivery feature into a support vector regressor (SVR) model with a Radial Basis Function (RBF) kernel (FT-SVR). The second one is based on an NN-based method [10] utilizing the BLSTM and attention mechanism with delivery, language use, and content features as input (FT-BLSTM).

3.3.1. Performance under the new task of seen type

The results for the new task of seen type are shown in Table 2, indicating under the low-resource setting, the proposed method outperforms the baselines by 5% in PCC. Also, both pre-training and meta-learning methods perform better than the FT-based methods.

Table 2. Comparison under the new task of seen type

Model	$\% \leq 0.5$	$\% \leq 1$	PCC(%)
FT-SVR[7]	78.3	95.4	65.5
FT-BLSTM[10]	80.1	96.3	67.3
MTL-finetune	82.1	96.5	72.1
Ours	**85.4**	**97.3**	**75.3**

3.3.2. Performance under the new task of unseen type

To further validate the adaptability of the proposed method, we test it under the new task of unseen type. Results are shown in Table 3. The proposed method performs much better than the MTL-finetune-based method, indicating its capability for fast adaptation.

Table 3. Comparison under the new task of unseen type

Model	$\% \leq 0.5$	$\% \leq 1$	PCC(%)
FT-SVR[7]	70.3	93.5	62.2
FT-BLSTM[10]	72.5	94.3	64.1
MTL-finetune	74.1	95.1	73.9
Ours	**78.5**	**96.7**	**78.7**

Table 4. PCC under different data sizes

Method	20	50	100	200
FT-SVR[7]	53.2	61.9	65.1	73.5
FT-BLSTM[10]	55.6	64.0	66.3	75.2
MTL-finetune	60.5	73.7	75.1	79.5
Ours	**63.5**	**78.4**	**79.2**	**80.1**

3.3.3. Performance under different adaptation data sizes

We investigate the relationship between the adaptation performance with different adaptation data sizes for the new task of unseen type (Opinion task). We split 200 data for training and 1000 for testing, and randomly sample a portion of training data for adaptation. The results are shown in Table 4. From the results, we can see only when the data size increases to about 200, the other baselines can achieve comparable results to the proposed method.

3.4. Ablation study

We validate the structure of the proposed method by removing the memory network (MN), which stores grading criteria, based on the dataset of Opinion task. Results are shown in Table 5. The deteriorating performance indicates the necessity of the MN unit for fast adaptation.

Table 5. Performance with no memory network

Model	$\% \leq 0.5$	$\% \leq 1$	PCC(%)
Ours (No MN)	76.3	94.7	75.2
Ours	**78.5**	**96.7**	**78.7**

4. CONCLUSION

In this paper, we attempt to tackle two main challenges in automatic spoken language assessment: adaptability and data scarcity. We propose a fast task-specific adaptation method with meta-learning. To improve adaptability under different grading criteria, we utilize a memory network to learn and store these criteria. Experimental results based on new tasks of seen and unseen types demonstrate its superiority to the traditional methods and MTL-finetune-based methods. In the future, we will investigate the possibility of fusing feature extraction and scoring modules for end-to-end optimization.

5. REFERENCES

[1] Klaus Zechner and Xiaoming Xi, "Towards automatic scoring of a test of spoken language with heterogeneous task types," in *Proceedings of the third workshop on innovative use of NLP for building educational applications*, 2008, pp. 98–106.

[2] Horacio Franco, Leonardo Neumeyer, Yoon Kim, and Orith Ronen, "Automatic pronunciation scoring for language instruction," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1997, vol. 2, pp. 1471–1474.

[3] Silke M Witt and Steve J Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] Wenping Hu, Yao Qian, and Frank K Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Interspeech*, 2013, pp. 1886–1890.

[5] Wenting Xiong, Keelan Evanini, Klaus Zechner, and Lei Chen, "Automated content scoring of spoken responses containing multiple parts with factual information," in *Speech and Language Technology in Education*, 2013.

[6] Klaus Zechner, Keelan Evanini, Su-Youn Yoon, Lawrence Davis, Xinhao Wang, Lei Chen, Chungmin Lee, and Chee Wee Leong, "Automated scoring of speaking items in an assessment for teachers of english as a foreign language," in *Proceedings of the ninth workshop on Innovative Use of NLP for Building Educational Applications*, 2014, pp. 134–142.

[7] Anastassia Loukina, Nitin Madnani, and Aoife Cahill, "Speech-and text-driven features for automated scoring of english speaking tasks," in *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, 2017, pp. 67–77.

[8] Andrey Malinin, Kate Knill, Anton Ragni, Yu Wang, and Mark JF Gales, "An attention based model for off-topic spontaneous spoken response detection: An initial study," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*. ISCA, 2017, pp. 144–149.

[9] Chong Min Lee, Su-Youn Yoon, Xihao Wang, Matthew Mulholland, Ikkyu Choi, and Keelan Evanini, "Off-topic spoken response detection using siamese convolutional neural networks.," in *INTERSPEECH*, 2017, pp. 1427–1431.

[10] Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian, "End-to-end neural network based automated speech scoring," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6234–6238.

[11] Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang, "A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 979–986.

[12] Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki, "Meta-adapter: Efficient cross-lingual adaptation with meta-learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7028–7032.

[13] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li, "Meta-learning for low-resource neural machine translation," *arXiv preprint arXiv:1808.08437*, 2018.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[15] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus, "End-to-end memory networks," *arXiv preprint arXiv:1503.08895*, 2015.

[16] Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang, "Neural approaches to automated speech scoring of monologue and dialogue responses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8112–8116.

[17] Suma Bhat and Su-Youn Yoon, "Automatic assessment of syntactic complexity for spontaneous speech scoring," *Speech Communication*, vol. 67, pp. 42–57, 2015.

[18] 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.

[19] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.