Multi-Stage Pre-training for Automated Chinese Essay Scoring

Wei Song[1], **Kai Zhang**[1], **Ruiji Fu**[2,3], **Lizhen Liu**[1], **Ting Liu**[4], **Miaomiao Cheng**[1]
[1]College of Information Engineering and Academy for Multidisciplinary Studies,
Capital Normal University, Beijing, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[3]iFLYTEK AI Research (Hebei), Langfang, China
[4]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
{wsong, kzhang, liz_liu7480, B365}@cnu.edu.cn,
rjfu@iflytek.com, tliu@ir.hit.edu.cn

Abstract

This paper proposes a pre-training based automated Chinese essay scoring method. The method involves three components: *weakly supervised pre-training, supervised cross-prompt fine-tuning* and *supervised target-prompt fine-tuning*. An essay scorer is first pre-trained on a large essay dataset covering diverse topics and with coarse ratings, i.e., *good* and *poor*, which are used as a kind of weak supervision. The pre-trained essay scorer would be further fine-tuned on previously rated essays from existing prompts, which have the same score range with the target prompt and provide extra supervision. At last, the scorer is fine-tuned on the target-prompt training data. The evaluation on four prompts shows that this method can improve a state-of-the-art neural essay scorer in terms of effectiveness and domain adaptation ability, while in-depth analysis also reveals its limitations.

1 Introduction

Automated essay scoring (AES) is an important educational application of natural language processing (NLP) (Page, 1966). AES aims to automatically judge the quality of student essays, which can reduce teachers' burden on essay scoring and provide fast feedback to students.

AES is usually viewed as a supervised learning problem. Traditionally, AES systems are based on hand-crafted surface-level features (Larkey, 1998; Attali and Burstein, 2006; Chen and He, 2013; Phandi et al., 2015). Recently, neural network based representation learning has been applied and achieved superior performance compared with traditional methods (Taghipour and Ng, 2016; Cummins et al., 2016; Alikaniotis et al., 2016; Dong and Zhang, 2016; Dong et al., 2017; Tay et al., 2018).

Most of the proposed methods, no matter the feature based or the representation learning based ones, work in an in-domain setting that is to train a scorer for a specific prompt based on a set of example essays for this prompt, and use this scorer to rate more essays from the same prompt. This manner usually requires many rated examples to get acceptable performance.

Although cross-domain transferable essay scoring has gained more attention (Phandi et al., 2015; Jin et al., 2018), the progress is still limited. The possible reason may be that the available corpora for essay scoring usually cover narrow topics on a small scale, and the topics, scoring criteria, and score ranges of different prompts often vary. Since an AES system should have the ability to appreciate or criticize essays, supervised pre-training is necessary. Intuitively, if a reader has read many rated essays from different prompts, she should be more experienced to judge the quality of an essay that responds to a new prompt. At least, she should require less guidance compared to a novice.

In this work, we empirically evaluate a pre-training based method for AES. Figure 1 illustrates the main framework. Our method has three components, each of which incorporates different level supervision. The first component is **weakly supervised pre-training**. An essay scorer is pre-trained based on a large scale essay corpus. The corpus covers diverse topics and is prompt-free. The essays are collected from the Web and have been rated by anonymous teachers. The essays' ratings are converted to binary coarse ratings: *good* and *poor* for the ease of weakly supervised pre-training. The second component is **supervised cross-prompt pre-training / fine-tuning**. This component aims to exploit the supervision from the training data of other prompts to pre-train or further fine-tune an essay scorer. The third component is **supervised target-prompt fine-tuning**. The pre-trained scorer would be fine-tuned on the training data for target prompts. Since human rat-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6723–6733,
November 16–20, 2020.

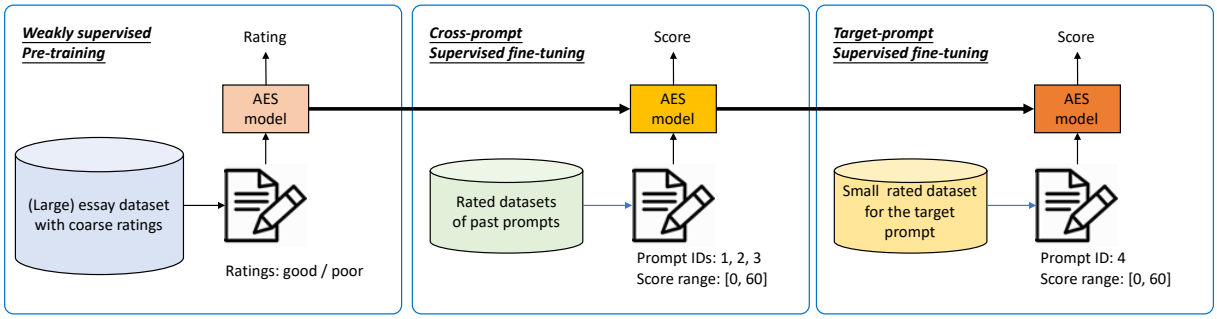


Figure 1: The proposed pre-training framework for automated essay scoring (AES).

ings are expensive to be collected, we expect the essay scorer depends on the target-prompt training data the less the better.

Although there are public available datasets in English such as the ASAP dataset.[1] These datasets usually cover only a few topics making it difficult to find datasets for pre-training and fine-tuning. As a result, we collect datasets and conduct experiments for automated Chinese essay scoring. We built a dataset with more than 85,000 essays written by junior and senior high school students for weakly supervised pre-training. We also collected nearly 4,000 essays in response to four prompts from senior high schools. These essays were carefully rated by teachers and are used for cross-prompt fine-tuning and evaluation.

Although the framework is straightforward, the evaluation demonstrates the effectiveness of the proposed method.

(1) **Higher performance in general**: The cooperation of the three components can improve the attentional recurrent convolutional neural network model (ARCNN) (Dong et al., 2017), which achieved the state-of-the-art result on the ASAP dataset. In average, the best pre-training enhanced ARCNN can achieve a 4.2% absolute improvement in QWK and 3.1% absolute improvement in Pearson coefficient compared with the ARCNN that is trained on the target-prompt training data only.

(2) **Better domain adaptation ability**: With both weakly pre-training and cross-prompt fine-tuning, our method can use 10% target-prompt training data (about 50 essays) to achieve 93.6% relative performance of the full model which is trained with 100% training data. Supervised cross-prompt fine-tuning is essential for domain adaptation though it is also expensive due to the requirement of human rated essays. With weakly pre-training only, our method can use half of the training data to achieve the same performance as the base scorer that is trained with 100% training data but without pre-training.

To the best of our knowledge, we are the first to investigate multi-stage pre-training based AES. We conduct careful analysis to gain more insights about how the method works and its limitations. Although our research focuses on Chinese, the results and observations should be useful for AES in other languages as well.

2 Related Work

AES is commonly viewed as a supervised learning problem with various feature templates (Larkey, 1998; Attali and Burstein, 2006; Chen and He, 2013; Phandi et al., 2015; Cummins et al., 2016; Song et al., 2017). These methods assume that essay quality correlates with surface-level features. The drawbacks of these methods include that the feature design and engineering are difficult and the semantic understanding of essays is limited.

Since 2016, neural network based AES systems become popular (Taghipour and Ng, 2016; Cummins et al., 2016; Alikaniotis et al., 2016; Dong and Zhang, 2016; Dong et al., 2017; Tay et al., 2018). These models obtained superior performance compared with traditional methods.

However, most of these systems are prompt-specific. New training data has to be annotated for training a new model for a new prompt.

Domain Adaptation for AES Phandi et al. (2015) proposed domain adaptation as a solution to adapt an AES system from one initial prompt to another prompt based on Bayesian linear ridge regression. Dong and Zhang (2016) demonstrated that the hierarchical CNN based model performs better in domain adaptation setting. Pilán et al. (2016); Xia et al. (2016) also attempted to incorpo-

[1] https://www.kaggle.com/c/asap-aes/

rate external knowledge for readability assessment. However, these methods mostly focused on domain adaptation from one domain to another but did not explore external resources.

Pre-training for AES Recently, pre-training language models (LM) becomes a trend (Devlin et al., 2019; Yang et al., 2019), which leads to the pre-training then fine-tuning mechanism and achieves great success in many NLP tasks.

For AES, Mim et al. (2019) proposed an unsupervised pre-training approach for evaluating the organization and argument strength of argumentative essays, where coherence modeling is used for pre-training. Rodriguez et al. (2019) attempted to apply BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019) for AES, but the results on ASAP are similar to the performance of a LSTM based scorer.

Howard and Ruder (2018) proposed the universal language model fine-tuning approach for text classification, including components such as general-domain LM pre-training, target domain LM fine-tuning and target task classifier fine-tuning. Gururangan et al. (2020) showed that task-adaptive pretraining can provide a large performance boost for ROBERTA across four domains and eight classification tasks. Motivated by previous works, this paper also adopts a multi-stage pre-training strategy by exploiting weak, distant and target oriented supervision for AES.

3 The Proposed Method

3.1 The ARCNN Model

Our base model is the attentional recurrent convolutional neural network model (ARCNN) (Dong et al., 2017), which is one of the state-of-the-art neural AES systems.

Sentence Representation A sequence of words $x = \{w_1, ..., w_N\}$ is modeled with a CNN encoder. The feature representation for the i-th word is

$$\mathbf{z}_i = f(\mathbf{W}_z \cdot [\mathbf{e}(w_i) : \mathbf{e}(w_{i+h_w-1})] + \mathbf{b}_z), \quad (1)$$

where we use $\texttt{tanh}$ as the activation function f, $\mathbf{e}(w_i) \in \mathbb{R}^d$ is the embedding of a word, h_w is the window size in the convolutional layer, $\mathbf{W}_z$ and $\mathbf{b}_z$ are weight matrix and bias vector.

Above the convolutional layer, attention pooling is employed to get the sentence representation $\mathbf{s}$,

$$\mathbf{s} = \sum \alpha_i \mathbf{z}_i, \quad (2)$$

where,

$$\alpha_i = \frac{e^{\mathbf{W}_\alpha \cdot \mathbf{m}_i}}{\sum e^{\mathbf{W}_\alpha \cdot \mathbf{m}_i}}, \mathbf{m}_i = \texttt{tanh}(\mathbf{W}_m \cdot \mathbf{z}_i + \mathbf{b}_m),$$

$\mathbf{W}_\alpha, \mathbf{W}_m, \mathbf{b}_m$ are parameter matrixes and bias vector for computing attentions.

Text Representation The sentence representations are modeled with a LSTM to get a sequence of hidden states $\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_S\}$, where S is the number of sentences. The hidden state of the j-th sentence is

$$\mathbf{h}_j = \texttt{LSTM}\left(\mathbf{s}_j, \mathbf{h}_{j-1}\right), \quad (3)$$

where $\mathbf{s}_j$ is the representation of the j-th sentence, and $\mathbf{h}_{j-1}$ is the hidden state of the previous step. Two LSTM encoders are applied in both directions and the bidirectional hidden representations are concatenated together to represent each sentence. The whole sequence could be represented as a fixed length vector $\mathbf{o} = \phi(\{\mathbf{h}_1, \cdots, \mathbf{h}_S\})$, where $\phi(\cdot)$ is a function to summarize hidden states. The attention mechanism are used as $\phi(\cdot)$ to get the text representation.

The Prediction Layer Finally, the rating of the essay is predicted according to

$$y = \texttt{sigmoid}(\mathbf{w}_y \cdot \mathbf{o} + \mathbf{b}_y), \quad (4)$$

where $\mathbf{w}_y$ and $\mathbf{b}_y$ are weight vector and bias vector.

3.2 Weakly Supervised Pre-training

We attempt to explore corpora with diverse topics and weak/distant quality judgements for pre-training a general essay scorer.

3.2.1 Data Collection

We collected essays from a website LeleKetang.[2] The essays were written by Chinese students in grade 7 to12. The corpus covers diverse topics and multiple genres, including *narrative, argumentative* and *prose* essays. The average number of sentences and Chinese characters are 30 and 779.

Each essay was rated by a teacher to indicate its quality before it was uploaded to the website. The ratings range from 1 to 4, indicating *poor*, *normal*, *good* and *excellent*. However, the ratings are imbalanced. Rating 3 and rating 1 are many more than rating 2 and rating 4. The corresponding statistics are shown in Table 1.

For pre-training, we combine rating 4 and 3 to represent ***good*** essays, view rating 1 as ***poor*** essays, and remove rating 2 to ensure that the good and poor essays could be distinguished.

[2] http://www.leleketang.com/zuowen/

Ratings ↓	#. Essays						Sum.
Grades →	7	8	9	10	11	12	
4	132	110	263	137	83	680	1405
3	5,510	5,337	5,754	4,744	1,684	4,079	27,108
2	1,886	1,556	1,273	1,122	733	897	7,467
1 (Poor)	15,229	13,550	10,445	5,696	6,107	5,995	57,022
4+3 (Good)	5,642	5,447	6,017	4,881	1,767	4,759	28,513

Table 1: The basic statistics of the dataset used for weakly pre-training. We combine rating 3 and rating 4 as *good* essays, and use rating 1 as *poor* essays. Rating 2 is not used in this work.

3.2.2 Pre-training the ARCNN Model

Formally, we have an essay dataset $E = \{(x, y)\}$, where $y \in \{0, 1\}$ indicates a *poor* or *good* essay.

We train the ARCNN model on the dataset E to distinguish *good* and *poor* essays. The learning objective is the sum of the negative cross-entropy over all training examples.

Since the collected ratings might be noisy and are converted to coarse binary ratings, we call it ***weakly supervised pre-training (WSP)***.

Parameters	Value
Embedding size	768
CNN window size	5
CNN filters	128
Dimension of LSTM hidden state	128
Batch size	32
Dropout (after embedding layer) ratio	0.5
Optimizer	Adam
Learning rate	0.0001

Table 2: Hyper-parameter values.

3.3 Supervised Fine-tuning

3.3.1 Supervised Target-Prompt Fine-tuning

The WSP model is just pre-trained on the coarse ratings so that its predictions are within the range of [0,1], which is different from the score ranges in real examinations. Moreover, the essays should be closely related to the prompts. As result, the model should be fine-tuned on the training data of target prompts.

Following Dong et al. (2017), the real scores are scaled to the range $[0, 1]$ for fine-tuning:

$$y_{scaled} = \frac{\hat{y} - min}{max - min}, \tag{5}$$

where $\hat{y}$ is the real score, min and max indicate the minimum and maximum scores in the training data. In evaluation phase, the predicted scores are rescaled to integer scores in the original score range.

The token representations are fixed during fine-tuning, which is the same as the pre-training. The other parameters would be fine-tuned. We call this strategy ***WSP-Finetune***.

3.3.2 Supervised Transfer Fine-tuning

If rated essays that are from other prompts are available, such data could be used to further train our weakly pre-trained model WSP before fine-tuning the model on target prompts. We just continue to fine-tune WSP on the available prompt-specific rated datasets. Since the rating knowledge learned from cross-prompt data would be transferred for scoring target-prompt essays, we call this strategy ***supervised transfer fine-tuning (Trans)***.

To be consistent with the score range of the target prompt, we only choose the essay datasets that have the same score range with the target prompt for supervised transfer fine-tuning. The main procedure is the same as described in Section 3.3.1. We put Trans before target-prompt fine-tuning so that the complete model is noted as ***WSP-Trans-Finetune***. Of course, Trans could be also used for pre-training if the weakly supervised pre-training data is not available, noted as ***Trans-Finetune***.

4 Evaluation

4.1 Model Parameter Settings

We use the tokenizer of BERT (Devlin et al., 2019) to get tokens and token embeddings. The vocabulary size is 21,128. The dimension of token embeddings is 768. The token embeddings are fixed during both pre-training and fine-tuning phases. We segment an essay into sentences by punctuation. The length limit of each sentence is set to 50. If the length of a sentence is longer than 50, it would be truncated and the remaining part is viewed as another sentence. The detail settings of hyper-parameters are listed in Table 2.

Pre-training dataset	P	R	F1	Acc.
Dev	0.72	0.71	0.71	0.75
Test	0.71	0.70	0.70	0.74

Table 3: Results of the pre-training task on the development and test data.

	# Essays	Avg. #sent.	Avg. #chars	Range
Set 1	964	24	819	0-60
Set 2	990	25	785	0-60
Set 3	866	25	781	0-60
Set 4	1,065	23	791	0-60

Table 4: Basic statistics of the target-prompt datasets.

4.2 Evaluation on Pre-training Task

4.2.1 Settings

Data We conducted experiments on the LeleKetang dataset. 80%, 10% and 10% of the dataset are used as the training, development and test data.

Evaluation metrics Since the coarse ratings are binary, we view pre-training as a classification problem. Macro precision(P), recall(R), F1-score (F1) and accuracy (Acc.) are used as evaluation metrics.

4.2.2 Results

Table 3 shows the experimental results on the development and test data of the pre-training dataset. The performance is moderate. The macro F1 score is about 0.74. This indicates that these essays are distinguishable on a certain degree.

Notice that the dataset covers diverse topics and different genres so that this task is not easy because different types of essays should be judged with different evaluation criteria. We also tried to incorporate genre and grade information in a multi-task learning setting for pre-training, but the results on the pre-training dataset and target prompts are not obviously better than using coarse ratings only. The acceptable results indicate that essays in different topics and genres should still share features that can indicate the quality of essays.

4.3 Evaluation on Target Prompts

4.3.1 Settings

Dataset We used four prompts which were previously used for writing test in college entrance examinations by two provinces in China, during 2012-2014. Each prompt is a short text describing an event, a quote, a fable or other background information (see Appendix A). We let students from several senior high schools write an essay according to their understandings of each prompt. The collected essays were scored by high school teachers. Each essay was scored by two teachers. The scores range from 0 to 60. If the difference between their scores is not bigger than 6 (10% of the score range), the average score would be the final score. Otherwise, a third teacher would participate in evaluation, and the average of two closest scores among the three would be the final score. This procedure is the same as the evaluation procedure in college entrance examinations. The collected essays are grouped according to prompts. The statistics of the datasets are shown in Table 4.

Evaluation Metrics We use the quadratic weighted Kappa (QWK) and Pearson coefficient score as evaluation metrics. QWK is widely adopted for evaluating AES, while Pearson coefficient could reflect ranking consistency.

We conducted 5-fold cross-validation. In each run, we used 60%, 20% and 20% of a dataset for each prompt as training data, development data and test data, respectively. The average performance would be reported.

Comparisons We compare the following systems. The first set of systems are previously proposed neural AES systems, including

- Taghipour and Ng (2016): This method uses CNN for word sequence modeling and LSTM for text level modeling. The text representation is obtained through mean of time pooling.
- Dong and Zhang (2016): This method uses a hierarchical CNN structure for modeling sentence and text representations.

The second set of systems are the variations of the proposed pre-training based AES series. All variations use ARCNN as the base model.

- ARCNN: The ARCNN model is trained only based on the target-prompt training data for each prompt.
- WSP: The ARCNN model is weakly pre-trained on the LeleKetang dataset and then directly used to predict target-prompt test data without fine-tuning.
- WSP-Finetune: The weakly supervised pre-trained model is further fine-tuned based on

Model	Set1		Set2		Set3		Set4		Average	
	QWK	Pears.	QWK	Pears.	QWK	Pears.	QWK	Pears.	QWK	Pears.
Dong and Zhang (2016)	0.710	0.754	0.517	0.574	0.286	0.364	0.450	0.513	0.491	0.551
Taghipour and Ng (2016)	0.779	0.789	0.569	0.626	0.392	0.459	0.559	0.586	0.574	0.615
ARCNN	0.842	0.856	0.574	0.625	0.355	0.441	0.563	0.602	0.584	0.631
Trans	0.775	0.853	0.541	0.598	0.319	0.375	0.453	0.487	0.522	0.578
Trans-Finetune	0.862	0.875	0.581	**0.645**	0.451	0.514	0.561	**0.608**	0.614	0.661
WSP	0.179	0.596	0.060	0.350	0.086	0.403	0.053	0.219	0.095	0.392
WSP-Finetune	0.862	0.872	0.580	0.623	0.465	0.500	0.564	0.607	0.618	0.651
WSP-Trans-Finetune	**0.863**	**0.877**	**0.586**	0.629	**0.495**	**0.534**	**0.567**	0.606	**0.628**	**0.662**

Table 5: QWK and Pearson coefficient scores on target-prompt test sets. All models are trained or fine-tuned using the full target-prompt training data.

the target-prompt training data and development data.

- Trans: Other prompt-specific training data is used to pre-trained a model. In experiments, for each target-prompt test data, we use the training data and development data of the other three prompts for scorer training and model selection.

- Trans-Finetune: This setting further fine-tunes the Trans model based on the target-prompt training data and development data.

- WSP-Trans-Finetune: The weakly supervised pre-trained model is fine-tuned on the cross-prompt data before being fine-tuned on the target-prompt data.

4.3.2 Overall Results

Table 5 shows the performance of the previous neural AES models and the variants of our proposed pre-training based models.

ARCNN obtains competitive performance compared with the other neural scorers. The results verify that ARCNN is an effective neural essay scorer and this is also the reason that we use it as the base model for pre-training.

Our final model WSP-Trans-Finetune achieves the best performance in average and outperforms ARCNN, which is trained and test on the datasets from the same prompts. The improved QWK score and Pearson coefficient score in average are 4.4% and 3.1%. The final model also outperforms Trans-Finetune and WSP-Finetune in most cases. This results verify that the multi-stage pre-training strategy is feasible and effective for AES in general.

One issue is that the performance gain across datasets is inconsistent. The improvement on Set3 is large, while the improvement on Set4 is relatively small.

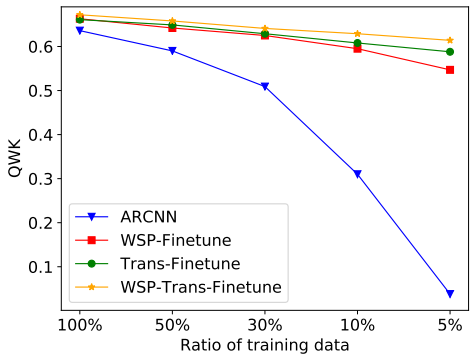
In addition, we can see that fine-tuning on the target-prompt training data is still essential. The performance of Trans decreases a lot without fine-tuning, while the WSP model is infeasible to be directly applied for scoring due to the different score ranges between pre-training and target-prompt data.

4.3.3 Analysis and Discussions

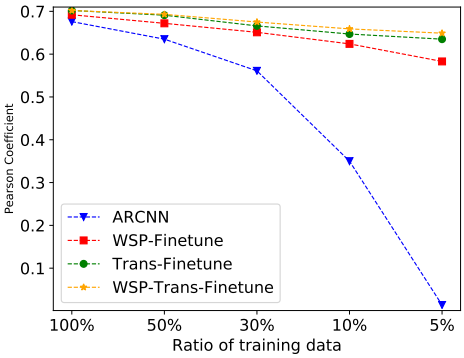
We provide more detail analysis and discussions from several aspects.

The effect of weakly supervised pre-training As shown in the last three rows in Table 5, when WSP is directly applied to score essays from target prompts, the QWK scores are very low. This is reasonable since the distribution of the coarse ratings in the pre-training dataset is far from the distribution of scores in target-prompt dataset. As a result, the differences between predicted scores and real scores are large, which lead to low QWK scores. However, the Pearson coefficient scores are not so low as QWK scores. This indicates that the weakly supervised pre-training can help capture some common indicators of the quality of essays without considering prompt specific information. After WSP is fine-tuned on the target prompts, WSP-Finetune obtains improvements on all four datasets.

The effect of supervised transfer pre-training Trans-Finetune pre-trains a model on narrow topics (3 prompts) but performs surprisingly well. Trans-Finetune may provide a kind of regularization to improve the generalization of the essay scorer. This explanation to the effectiveness of pre-training is well accepted (Erhan et al., 2010). Moreover, more training data from the same score range also helps shape the real distribution of scores and avoids overfitting to the distribution in the target-prompt training data.



(a) QWK scores.



(b) Pearson coefficient scores.

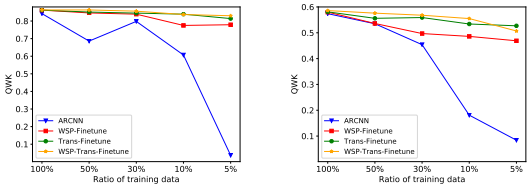
Figure 2: Average performance of training/fine-tuning with different ratio of target-prompt training data.

The combination of Trans and WSP WSP-Trans-Finetune achieves the best performance but its advantage compared with Trans-Finetune and WSP-Finetune is not very obvious, indicating that Trans and WSP benefit each other but also play similar roles.

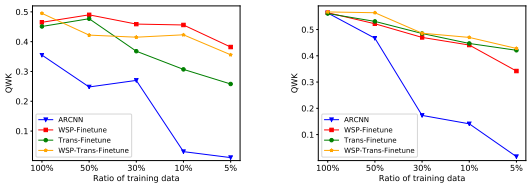
On one hand, both Trans and WSP can play a role as regularization. Because the topics are still narrow for cross-prompt pre-training so that new bias might be brought in, while WSP can help alleviate such an effect. On the other hand, WSP is trained based on coarse binary ratings. Trans can help WSP adapt the prediction distribution towards the score range of the target prompts.

Can pre-training reduce the requirement of target-prompt training data? This is a key question for this research. To answer it, we use different ratio of target-prompt training data to train ARCNN and fine-tune the pre-trained models. We sampled these subsets according to the score distribution of the whole dataset for each prompt.

Figure 2 shows the average QWK and Pearson



(a) Set1. (b) Set2.



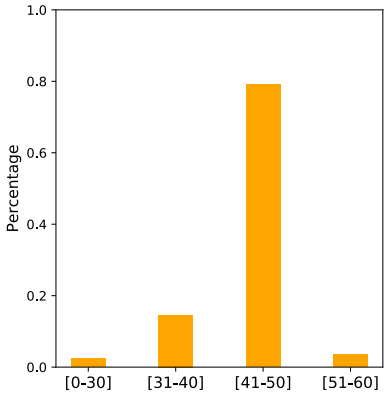
(c) Set3. (d) Set4.

Figure 3: The QWK scores with different ratio of target-prompt training data over four prompt datasets.

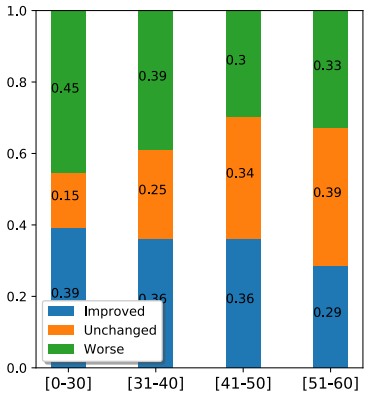
coefficient scores over four prompts with different ratio of training data. We can see that when the size of training data decreases, the performance of ARCNN drops sharply. In contrast, all three pre-trained models, WSP-Finetune, Trans-Finetune and WSP-Trans-Finetune, achieve very consistent performance even when the ratio of used training data is small. For example, in average, WSP-Finetune can use 50% target-prompt training data to obtain similar performance compared with ARCNN trained with all training data, and use 10% target-prompt training data to obtain 93.6% performance of ARCNN. Trans-Finetune and WSP-Trans-Finetune perform even better than WSP-Finetune. The cross-prompt supervised transfer fine-tuning (Trans) is useful for domain adaptation.

Figure 3 shows the QWK scores with different ratio of target-prompt training data across four prompts in detail. We can see that the trends on four datasets are generally consistent with the average performance. The pre-training based models outperform ARCNN with a large margin when the ratio of target-prompt training data is small. WSP-Trans-Finetune performs best on 3 datasets, while WSP-Finetune performs best on 1 dataset. Trans-Finetune obtains close performance compared with WSP-Trans-Finetune.

On one hand, these observations are encouraging. It means that if we have high quality rated cross-prompt essays, the supervised transfer pre-training can help a lot for domain adaptation. But such datasets are still expensive and large scale such datasets might be not always available. Even so,



(a) Distribution of essays from different ranges.



(b) The effect of WSP-Trans-Finetune compared with ARCNN.

Figure 4: The effect of WSP-Trans-Finetue on essays from different ranges compared with ARCNN.

the weak supervision through coarse ratings can also make an impact on domain adaptation.

On the other hand, the effects of different pre-training strategies are different at different datasets. This indicates that the effects of pre-training may be also related to the properties of target prompts. Moreover, we observe that in some cases (e.g, Set1 and Set3) using fewer training data (e.g., 30%) performs better than using more training data (e.g., 50%). This may relate to the representativeness of selected subsets of essays for training.

How does pre-training affect essays from different score ranges compared with ARCNN? We divide all the essays from four datasets into four ranges according to their real scores. The distribution of scores is shown in Figure 4(a). We can see that the essay scores are concentrated in range [40-50].

We analyze the WSP-Trans-Finetune model. We define *improvement* here as reducing the differences between the predicted and real scores compared with ARCNN. Figure 4(b) shows the results. We can see that the pre-training improves the scoring ability for essays from range [40-50]. So the general performance of WSP-Trans-Finetune is good. The essays from this range are at intermediate level, written in the common way. The pre-training models may help find subtle distinctions in style to distinguish them better.

	Set1	Set2	Set3	Set4
mean	0.422	0.489	0.435	0.445
median	0.420	0.492	0.434	0.445
quartile deviation	0.104	0.107	**0.123**	0.115
coefficient of variation	0.180	0.156	**0.201**	0.183

Table 6: Some statistics of Jensen-Shannon divergence between topic vectors of essays on four datasets. Each essay is represented with a topic distribution vector inferred by a LDA model.

However, pre-training hurts the performance in other ranges, although the number of essays in these ranges is small. The reasons might be as follows. High score prediction is a challenge for AES, because the training examples are less than other ranges and some high score essays were written in unique ways. Essays in the range [0-40] often involve off-topic essays. The pre-training models could not help much in these cases, because they can not help capture topic information very well.

Why the performance gain is inconsistent across prompts? We observe that the effects of pre-training vary across prompts, e.g., the performance gain in Set3 and Set4 is quite different.

Qualitatively, we speculate the inconsistence is related to the distinct properties of the prompts. For example, the prompt 3 has a semi-topic setting: writing an essay to discuss "____ *to know*", where the underline part should be filled in by students. So the students discussed this from a variety of angles. In this case, the importance of target-prompt examples might be weakened and pre-training plays an important role. The prompt 4 asked students to imagine a situation if we would have an intelligent chip which knows all kinds of knowledge. In this case, a good sense of imagination and creativity may become a scoring dimension to human raters. But this dimension is difficult to be captured by AES models.

We try to find quantitative evidence to support our speculations. We analyze the topical diversity of essays within each prompt. We train a LDA model with 200 topics on the pre-training dataset

and infer the topic distribution of each essay in four prompt datasets. We then compute the Jensen-Shannon divergence between every pair of essays. Table 6 shows some statistics of these values. Unfortunately, we do not find obvious regularity except that the essays from prompt 3 cover more diverse topics compared with other prompts according to the quartile deviation and the coefficient of variation. We leave the investigation of the correlation between datasets' properties and scoring performance as future work.

5 Conclusion

In this paper, we presented a pre-training based approach to automated Chinese essay scoring. Our method investigates multi-stage pre-training and incorporates multi-level supervision, including the weak supervision from large scale coarse ratings, the supervision from rated essays from other prompts and the target-prompt training data.

The experimental results show that the pre-training based approach is effective for AES in terms of both effectiveness and domain adaptation ability. We carefully analyze the effects of each component and find that: multi-stage pre-training improves the base model in general; the domain adaptation ability can be consistently improved; target-prompt fine-tuning is still indispensable but the required amount of training data can be largely reduced; weakly supervised pre-training and supervised transfer fine-tuning are both helpful.

We also observe some phenomena but do not have good explanations. For example, the performance gain across prompts is inconsistent. When the pre-trained scorer can work best should be further studied. We suggest that the prompts' properties should be investigated more for applying AES.

The proposed method has a limitation that it pays more attention to the score range that most essays are from, and may hurt the performance in other ranges. Another limitation of the method is the dependence on pre-training dataset. The pre-training dataset used in this paper is still small compared with the data used for pre-training language models. Larger pre-training dataset with supervised labels or self-supervised learning strategies could be explored. Moreover, we are interested in understanding what features or traits of essays are captured by the deep models for scoring. We plan to investigate these in future.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 61876113, 61876112), Beijing Natural Science Foundation (No. 4192017) and Capital Building for Sci-Tech Innovation-Fundamental Scientific Research Funds. Lizhen Liu is the corresponding author.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of ACL 2016*, pages 715–725.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of EMNLP 2013*, pages 1741–1752.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *Proceedings of ACL 2016*, pages 789–799.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of EMNLP 2016*, pages 1072–1077.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162.

Dumitru Erhan, Aaron C. Courville, Yoshua Bengio, and Pascal Vincent. 2010. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

Leah S Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of SIGIR 1998*, pages 90–95.

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2019. Unsupervised learning of discourse-aware text representation for essay scoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 378–385, Florence, Italy. Association for Computational Linguistics.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of EMNLP 2015*, pages 431–439.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv: Computation and Language*.

Wei Song, Dong Wang, Ruiji Fu, Lizhen Liu, Ting Liu, and Guoping Hu. 2017. Discourse mode identification in essays. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 112–122.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

A Appendix

The general contents of four prompts are listed below. The students were asked to write an essay no less than 800 Chinese characters according to the prompt. There are no restrictions on title and genre.

Prompt 1: 尚先生把手机落在出租车上。他随后拨打那部手机，对方接听后立即挂断。他又发短信表示，愿意出2000元“买”回手机。一小时后，尚先生收到回复，对方要归还手机。捡到手机的人是一位年轻人。尚先生要酬谢他，但对方交还手机后就转身离去了。当天晚上，记者联系到那位年轻人，年轻人说：“我本来无意归还，但看到手机里的照片和信息，发现机主刚刚给芦山地震灾区汇去一大笔捐款，很受感动。我不能见利忘义，不能用贪心对待爱心。我也要像尚先生那样多一些真诚和友善。”

Translation of prompt 1: Mr. Shang lost his phone on the taxi. He then dialed the phone, but the other party hung up after connecting. Mr. Shang sent a message, expressing his willingness to spend 2,000 yuan to “buy” the phone back. An hour later, the other party replied and was willing to return the phone. It was a young man who picked up the phone. Mr. Shang wanted to appreciate him with money, but the young man refused. The young man said that he had no intention of returning the phone at first, but when he saw the photos and messages in the phone, he noticed that Mr. Shang just made a substantial donation to the earthquake-stricken area. He was moved and decided to return it for love and forgot the covetous thoughts.

Prompt 2: 两条小鱼一起游泳，遇到一条老鱼从另一方向游来，老鱼向他们点点头，说:“早上好，孩子们，水怎么样?”两条小鱼一怔，接着往前游。游了一会儿，其中一条小鱼看了另一条小鱼一眼，忍不住说:“水到底是什么东西？”看来，有些最常见而又不可或

缺的东西，恰恰最容易被我们忽视;有些看似简单的事情，却能够引发我们深入思考?

Translation of prompt 2: Two little fishes swam together and encountered an old fish. The old fish nodded to them and said, "Good morning, boys, how is the water?" The little fishes were puzzled and continued to swim. After a while, one of the little fish glanced at the other one and asked, "What is the water?" It seems that some common but indispensable things are often ignored; some seemingly simple things can give us a deep thought.

Prompt 3: 中国自古有“学而知之”的说法，这里的“学”，通常被理解为从师学习。韩愈就说过：“人非生而知之者，孰能无惑?惑而不从师，其为惑也，终不解矣。”随着时代的发展，我们获取知识、掌握技能或懂得道理的途径日趋多元。请结合你的心得和体验，在“____而知之”中的横线处填入一字，构成题目，写一篇文章，不能以“学而知之”为题。

Translation of prompt 3: There has been a saying in China since ancient times that "learning to know". Learning here is usually understood as learning from a teacher. Han Yu once said that "no one is born to know, and everyone has confusions. If you have confusions but do not learn from a teacher, the confusions would be always there." With the development of the times, the ways we acquire knowledge or master skills are becoming more diverse. According to your experience, fill in a word (except *learning*) above the underline in ____ *to know*, and write an essay.

Prompt 4: 也许将来有这么一天，我们发明了一种智慧芯片，有了它，任何人都能古今中外无一不知，天文地理无所不晓。比如说，你在心里默念一声“物理”，人类有史以来有关物理的一切公式、定律便纷纷浮现出来，比老师讲的还多，比书本印的还全。你逛秦淮河时，脱口一句“旧时王谢堂前燕”，旁边卖雪糕的老大娘就接茬说“飞入寻常百姓家”，还慈祥的告诉你，这首诗的作者是刘禹锡，这时一个金发碧眼的小女孩说，诗名《乌衣巷》出自唐诗，这将是怎样的情形呀！读了以上材料，你有怎样的联想或思考？请就此写一篇文章。

Translation of prompt 4: Perhaps one day in the future, we would have invented an intelligent chip. With this chip, anyone can know everything from ancient to modern times, from astronomy to geography. For example, if you meditate on *physics* in your mind, all formulas and laws related to physics in the history of mankind have emerged, more than the teachers have taught, more than the books have told. When you visit the Qinhuai River, you blurt out a sentence of a poem. An old lady who is selling ice cream would say the next sentence and kindly tells you who the author of this poem is. At the same time, a blond little girl from another country tells that the poem is from Tang poetry. How amazing it is. Based on the above materials, please write an essay about your associations and imagination.