Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring

Jiaqi Lun,[1] **Jia Zhu,**[1,3,*] **Yong Tang,**[1] **Min Yang**[2]
[1]School of Computer, South China Normal University, Guangzhou, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]Guangzhou Key Laboratory of Big Data and Intelligent Education
{jiaqilun, jzhu, ytang}@m.scnu.edu.cn, min.yang@siat.ac.cn

Abstract

Automatic short answer scoring (ASAS) is a research subject of intelligent education, which is a hot field of natural language understanding. Many experiments have confirmed that the ASAS system is not good enough, because its performance is limited by the training data. Focusing on the problem, we propose MDA-ASAS, **m**ultiple **d**ata **a**ugmentation strategies for improving performance on **a**utomatic **s**hort **a**nswer **s**coring. MDA-ASAS is designed to learn language representation enhanced by data augmentation strategies, which includes back-translation, correct answer as reference answer, and swap content. We argue that external knowledge has a profound impact on the ASAS process. Meanwhile, the Bidirectional Encoder Representations from Transformers (BERT) model has been shown to be effective for improving many natural language processing tasks, which acquires more semantic, grammatical and other features in large amounts of unsupervised data, and actually adds external knowledge. Combining with the latest BERT model, our experimental results on the ASAS dataset show that MDA-ASAS brings a significant gain over state-of-art. We also perform extensive ablation studies and suggest parameters for practical use.

Introduction

In recent years, the number of online educational applications has been growing rapidly, consisting of intelligent tutoring systems, e-learning environments, distance education, and massive open online courses. Automatically scoring short student answers is important for building intelligent tutoring systems. In general, computer-aided assessment systems are particularly useful because scoring by humans can become monotonous and tedious (Kumar et al. 2017). Automatic scoring systems can help teachers save lots of time from duplication of marking student's homework. Formally, the problem of automatic scoring short answer is defined as scoring the question in the context of the student answer and its reference answer. Table 1 shows an example of a short answer scoring task.

*Corresponding author is Jia Zhu.

In general, traditional methods and neural network methods are employed in the ASAS task. Traditional methods are driven by handcrafted features, such as lexical similarity features (Dzikovska, Nielsen, and Brew 2012), clear criteria (Siddiqi, Harrison, and Siddiqi 2010), and graph alignment features (Mohler, Bunescu, and Mihalcea 2011). Neural network methods are driven by the rapid development of deep learning techniques (Mueller and Thyagarajan 2016; Zhao et al. 2017). Recently, Saha et al. (Saha et al. 2019) have used InferSent (Conneau et al. 2017) and neural domain adaptation to obtain state-of-art results in the ASAS task. Deep learning has proven to be effective in long text NLP tasks. Due to the lack of information in the short sentence of the ASAS corpus, it seems not good enough in the ASAS task. The urge to obtain more information is the key to the current problem.

In fact, data augmentation is very popular in the research areas of vision (Krizhevsky, Sutskever, and Hinton 2012) and speech (Ko et al. 2015). However, it is rarely applied in the NLP task. In the past three years, many scholars have explored some methods to augment data, which have also proven to be effective in the ASAS task. General methods of text data augmentation are to replace words with their synonyms selected from a handcrafted ontology such as WordNet (Zhang, Zhao, and LeCun 2015) or word similarity calculation (Wang and Yang 2015). However, these methods are done at the word level, which is easy to lose the semantic information of the entire sentence.

In this paper, we propose MDA-ASAS, multiple data augmentation strategies for improving performance on automatic short answer scoring. It is designed to learn language representation enhanced by data augmentation strategies, which includes three data augmentation strategies. In the first strategy, we propose back-translation to augment the ASAS training data. The second strategy, we hypothesize that the student answers which received correct scores by the teacher are equivalent to teacher-provided reference answers. The third strategy, we propose a swapping content method between original data and twin data to achieve the best performance of data augmentation. Meanwhile, the BERT (Devlin et al. 2018) model has achieved excellent results in question-answer and natural language inference.

Question	You used several methods to separate and identify the substances in mock rocks. How did you know the crystals were salt?
Reference Answer	The water was evaporated, leaving the salt.
Student#1	We poured just the water into another thing and let the water evaporate.
Student#2	By looking very closely at it.

Table 1: An illustrative example showing question, reference answer, and student answers (Student#1 and Student#2) from a science course.

Therefore, our MDA-ASAS combines with the latest BERT model, which can bring a significant gain.

We demonstrate that MDA-ASAS combines with the latest BERT model, which leverages the power of external knowledge to enhance the ASAS task. The paper makes the following contributions:

- We propose three data augmentation strategies to enrich the training dataset and can help get a better representation of the sentences.
- Our MDA-ASAS combines with the latest fine-tuned BERT model for the ASAS task, which can bring a significant gain.
- We also make extensive ablation studies and suggest parameters for practical use in the ASAS task.

The rest of the structure of this paper is constructed as follows: after reviewing related work in section 2, we present three data augmentation strategies in section 3. Section 4 reports the experiments. And section 5 concludes the paper.

Related Work

In this section, the prior work is divided into three relevant research areas, automatic short answer scoring, transfer learning in ASAS and data augmentation in NLP.

Automatic Short Answer Scoring

Traditional methods utilize handcrafted features, such as lexical similarity features (Dzikovska, Nielsen, and Brew 2012), a clear criteria (Siddiqi, Harrison, and Siddiqi 2010), graph alignment features (Mohler, Bunescu, and Mihalcea 2011), n-gram features (Heilman et al. 2013), softcardinality text overlap features (Jimenez, Becerra, and Cic-Ipn 2013), averaged word vector text similarity features (Sultan, Salazar, and Sumner 2016) and other shallow lexical features (Ott et al. 2013).

More recently, deep learning approaches have been utilized for the automatic short answer scoring task. Mueller et al. (Mueller and Thyagarajan 2016) proposed a siamese adaptation of the LSTM network for labeled data comprised of pairs of variable-length sequences. Zhao et al. (Zhao et al. 2017) proposed an efficient memory networks-powered automated scoring model. Riordan et al. (Riordan et al. 2017) explored simple LSTM and CNN-based architectures for short answer scoring. Kumar et al. (Kumar et al. 2017) proposed a method involving Siamese biLSTMs, a novel pooling layer based on the Sinkhorn distance between LSTM state sequences, and a support vector ordinal output layer.

Transfer Learning in ASAS

Transfer learning has been shown to be effective for improving many natural language processing tasks. In general, transfer learning actually adds external knowledge because most of their models are learned using large amounts of unsupervised data. The short answer scoring problem is often modeled as a classification task. Thus, we can use the transfer learning downstream task to score the short answer. InferSent (Conneau et al. 2017) used a max pooled bidirectional LSTM network to learn universal sentence embeddings from the MultiNLI corpus (Williams, Nangia, and Bowman 2017). These embeddings have been employed as features in conjunction with hand-crafted features by Saha et al. (Saha et al. 2018) for ASAS. Hassan et al. (Hassan, Fahmy, and El-Ramly 2018) proposed a supervised learning approach for short answer automatic scoring based on paragraph embeddings, which included Word2Vec (Pennington, Socher, and Manning 2014), GloVe (Pennington, Socher, and Manning 2014), Fasttext (Joulin et al. 2016) and Elmo (Peters et al. 2018). Liu et al. (Liu, Xu, and Zhao 2019) used sentence embeddings by pre-trained BERT (Devlin et al. 2018) model to score automaticly.

Data Augmentation in NLP

Text data augmentation has been extensively studied in natural language processing. Prior work has explored using paraphrasing for data augmentation on NLP tasks. Zhang et al. (Zhang, Zhao, and LeCun 2015) augmented their data by swapping out words for synonyms from WordNet (Miller 1995). Wang and Yang (Wang and Yang 2015) used a similar strategy, but identified similar words and phrases based on cosine distance between vector space embeddings. Yu et al. (Yu et al. 2018) generated new data by translating sentences into French and back into English.

Jia and Liang (Jia and Liang 2016) proposed grammar induction to augment the training data. Silfverberg et al. (Silfverberg et al. 2017) proposed task-specific heuristic rules to generate new data. Bergmanis et al. (Bergmanis et al. 2017) proposed neural decoders of autoencoders to augment the training data. Xia et al. (Xia et al. 2017) proposed encoder-decoder models to augment the training data. Kobayashi et al. (Kobayashi 2018) proposed contextual augmentation with a bi-directional RNN language model to augment the training data. Wu et al. (Wu et al. 2019) proposed conditional BERT (Devlin et al. 2018) contextual augmentation to augment the training data. Recently, research has proposed some easy data augmentation techniques on text classification tasks (Wei and Zou 2019).

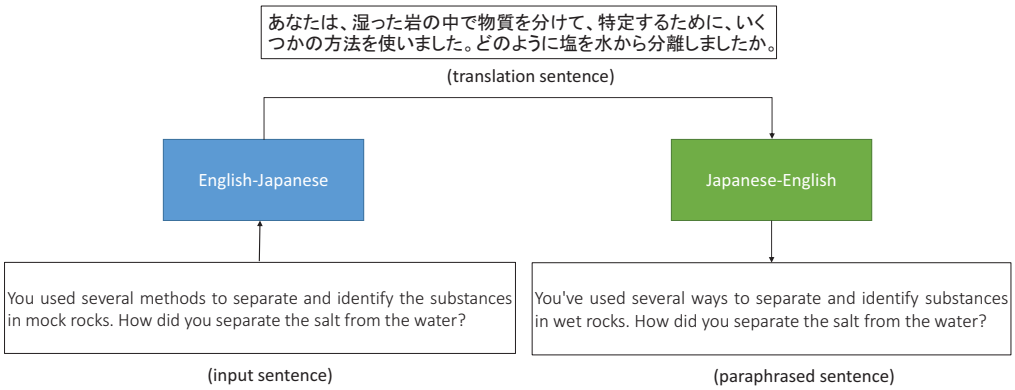

Figure 1: An illustration of the whole procedure of back-translation with Japanese as a key language.

Multiple Data Augmentation Strategies

We propose MDA-ASAS, multiple data augmentation strategies for improving performance on automatic short answer scoring. We discuss three data augmentation strategies for ASAS, including back-translation, correct answer as reference answer, and swap content.

Back-translation

Machine learning and deep learning have achieved high accuracy on the NLP tasks, but high performance often depends on the size and quality of training data, which is often tedious to collect. Since training data constrain the ASAS system performance, we can train it with much more data. Therefore, we combine our ASAS model with a simple data augmentation strategy to enrich the training data. The easy strategy idea is to use back-translation.

We observe prior work has explored using paraphrasing for data augmentation on NLP tasks. Zhang et al. (Zhang, Zhao, and LeCun 2015) augmented their data by swapping out words for synonyms from WordNet (Miller 1995). Wang and Yang (Wang and Yang 2015) used a similar strategy, but identified similar words and phrases based on cosine distance between vector space embeddings.

Previous data augmentation work is done at the word level, which it is difficult to keep the semantic information of the entire sentence. However, we use the back-translation for data augmentation, which is to translate the entire sentence and belongs to a sentence-level data augmentation, so that the whole sentence grammar, semantics, context, and other information can be retained intact.

Baidu translation is a popular translation tool in China, and the same translation tool as Google translate. Baidu translation currently supports the translation of 28 languages. Baidu translation has made significant breakthroughs in the acquisition of massive translation knowledge, translation model, multilingual translation technology, and other aspects, and responded to massive, complex, and diverse translation requests on the Internet in a timely and accurate manner. It has developed an online translation system that integrates deep learning with multiple mainstream translation models, leading the industry.

We develop a translation tool on Baidu translation API [1], which translate English to Japanese (or any other language), and then translate Japanese back to English, to obtain paraphrases of texts. This approach helps automatically increase the amount of training data for broadly any language-based tasks, including the ASAS task that we are interested in. With more data, we expect to generalize our model better and make our model more robust. The augmentation process is illustrated in Figure 1 with Japanese as a key language.

Correct Answer as Reference Answer

Our dataset is less than 5000 (reference answer, student answer) training pairs, which may not fully sustain the training of deep learning methods. To mitigate this problem, we propose a data augmentation strategy for short answer scoring by making use of correct student answers.

In this paper, we assume that the correct student answer is another form of reference answer to augment our training dataset. Here, we describe our data augmentation strategy. In our dataset, according to the teacher's analysis, a considerable number of student answers are given correct scores. Because there are many students whose answer label is correct, we hypothesize that the student answers which received correct scores by the teacher are equivalent to teacher-provided reference answers. As shown in Figure 2, the table on the left is an example of original data containing one question, one reference answer, and three student answers. Through the data augmentation strategy, 1 out of 3 students received correct scores $(1 < 3)$, then we can generate $1 \times (3-1) = 2$ new (reference answer, student answer) training pairs. Finally, we can get the data-enhanced ASAS data on the right table in Figure 2, which includes one question, two reference answers, and five student answers. For testing, we only use the reference answer provided by the teacher.

Swap Content

Our original training dataset contains 135 questions (q), 135 reference answers (r), and 4969 student answers (a). The

[1] api.fanyi.baidu.com

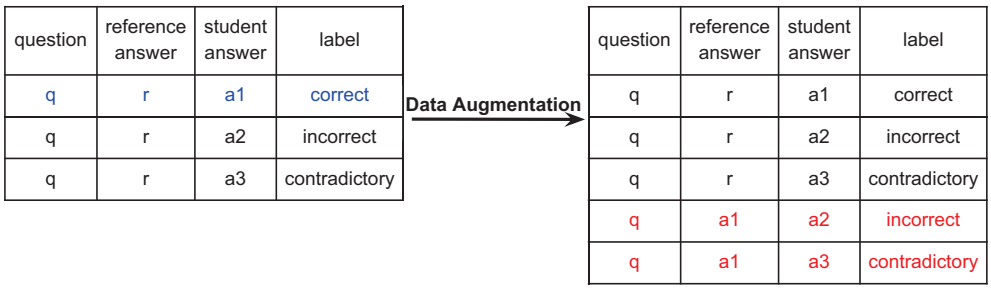

Figure 2: Firstly, the table on the left is an example of the original ASAS dataset, which contains a question, a reference answer, and three student answers. Secondly, through the correct answer as reference answer data augmentation strategy, 1 out of 3 students received correct scores $(1 < 3)$, then we can generate $1 \times (3 - 1) = 2$ new (reference answer, student answer) training pairs. So we can see that the red-marked data in the table on the right is enhanced data. Finally, the ASAS dataset includes one question, two reference answers, and five student answers.

original data is generated by the back-translation data augmentation strategy, which produces the back-translated data of the original data. We call this twin data, and it also includes 135 questions (q′), 135 reference answers (r′), and 4969 student answers (a′). The original data and the twin data have the same number of questions, reference answers, and student answers. For the data of the twin dataset, we carefully examine the original dataset with the naked eye and find that the meaning of each sentence corresponding to the original dataset is basically the same. Therefore, the data of the twin dataset is reliable.

We propose a third simple strategy of data augmentation, which is to randomly swap questions, reference answers, or student answers in the two datasets. We make sure that these two columns swap one to one. The augmentation process is illustrated in Figure 3 with two datasets.

The data augmentation process for randomly swapping two columns is as follows: First, we have two datasets, the original data, and the twin data. The original data includes questions, reference answers, student answers (q, r, a), and the twin data also includes questions, reference answers, and student answers (q′, r′, a′). Then the original data's question, reference answer, student answer, and twin data's question, reference answer, student answer corresponding column are swapped. Finally, through the above strategy, our training dataset can get eight different combinations, which are [(q, r, a), (q′, r′, a′)], [(q′, r, a), (q, r′, a′)], [(q, r′, a), (q′, r, a′)], [(q, r, a′), (q′, r′, a)].

Experiments

To evaluate the effectiveness of our models, we make several comparative experiments. In the following, we will introduce the dataset we used, our experiment settings, and experimental results in order.

Dataset

SemEval-2013 dataset: This dataset is SemEval-2013 Shared Task 7 dataset, which is a part of the "The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge" in the Semantic Evaluation workshop in 2013 (Dzikovska et al. 2013)[2]. We use the SciEntsBank corpus of SemEval-2013 dataset, which contains reference answers and student answers for 197 questions in 15 different science domains. Each of the three categorized subtasks has three different test sets have three different test sets, including Unseen Answers (UA), Unseen Questions (UQ) and Unseen Domains (UD). The three classification subtasks consist of 1) 2-way classification into correct and incorrect classes, 2) 3-way classification into correct, incorrect and contradictory classes, 3) 5-way classification into correct, partially correct, contradictory, irrelevant and non domain classes. The statistics for questions and student answers in the different test sets are shown in Table 2. For the dataset, the results are reported in terms of accuracy (Acc), macro-averaged F1 (M-F1) and weighted-average F1 (W-F1).

Experiment Settings

For the experiment settings subsection, we will introduce the details of fine-tuned BERT model and the details of implementation in the experiment.

Fine-tuned BERT Model in ASAS The BERT model has achieved excellent results in question-answer and natural language inference (NLI). In this paper, we argue that the BERT model, with its pre-training on a huge dataset and the powerful architecture for learning complex features, can further boost the performance of automatic short answer scoring. In this paper, we focus on using the fine-tuned BERT model to obtain sentence features for the ASAS task.

In order to fine-tune BERT$_{BASE}$, we follow the same approach of Devlin et al. The prediction task in ASAS is the task of assessing and scoring a student answer in comparison to a reference answer for the given question. The (reference answer, student answer) input is represented as a packed sequence with the reference answer assigned the A embedding and the student answer assigned the B embedding. As illustrated in Figure 4 in the paper, the first token of every sequence is the special classification embedding ([CLS]). The

[2]https://www.cs.york.ac.uk/semeval-2013/task7/data/uploads/datasets/

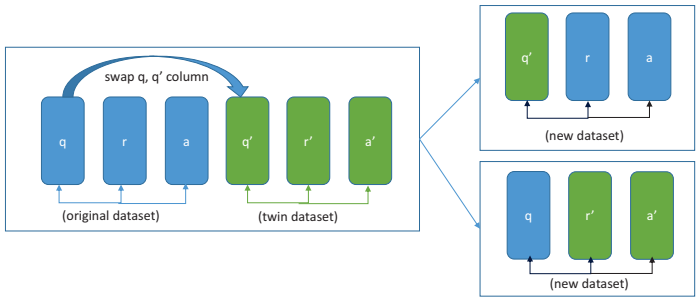

Figure 3: An example is given to illustrate the proposed data augmentation strategy, showing that the q column of original data is swapped the q′ column of twin data. So we generated two new datasets.

Data set	Question	Student answer
Training	135	4969
UA	135	540
UQ	15	733
UD	46	4562
Total	331	10804

Table 2: Question distribution and student answer distribution in the dataset.

final hidden state corresponding to this token is used as the aggregate sequence representation for the ASAS task. At the same time, [CLS] represents sentence features.

In the experiment, to obtain a fixed-dimensional pooled representation of the input reference answer and student answer, we take the final hidden state (i.e., the output of the Transformer) for the first token in the input, which by construction corresponds to the special ([CLS]) word embedding. We assume that this vector is $C \in R^H$. During the fine-tuning process, the new parameters are added for a classification layer $W \in R^{K \times H}$, where K is the number of classifier labels. The score probabilities $S \in R^K$ are computed with a standard softmax, $S = softmax(CW^T)$. All of the parameters of BERT and W are fine-tuned jointly to maximize the log-probability of the correct score. We leverage the power of BERT by using the BERT$_{BASE}$ model implemented HuggingFace [3].

Implementation Details We implement our experiments in PyTorch on an NVIDIA Corporation GM200. For the BERT model, we present a series of experiments using the Huggingface Pytorch BERT implementation for ASAS. And we use the BERT$_{BASE}$, which has 110 million parameters with L=12, H=768, and A=12. For the text encoder, the maximum length of the answers is set to 90 words. Finally, we fine-tune the BERT model with MDA-ASAS for ASAS to get the feature vector into softmax.

[3]https://github.com/huggingface/pytorch-pretrained-BERT

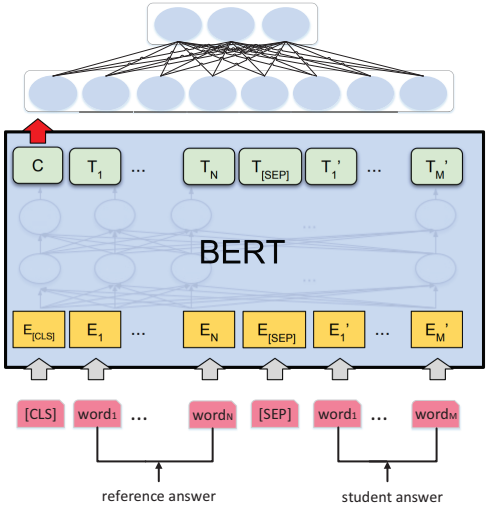

Figure 4: The BERT model in the ASAS task.

Results

In this part, we analyze the results of our experiments in detail and prove the effectiveness of our proposed MDA-ASAS. First, we prove the effectiveness of data augmentation. Then, we show a detailed ablation study to understand the effectiveness of MDA-ASAS better. Finally, we compare MDA-ASAS with the fine-tuned BERT model against previous state-of-art models on the dataset.

Effect of the Data Augmentation Strategies To prove the effect of data augmentation strategies, we conduct different experiments with three data augmentation strategies in the training data. Table 3 shows macro-averaged-F1 and weighted-average-F1 with MDA-ASAS on 2-way, 3-way, and 5-way of UA test sets. As described in the multiple data augmentation strategies section, we perform three methods of data augmentation on our original data training set, which includes back-translation, correct answer as reference answer, and swap content. Therefore, our training datasets consist of the original dataset (A-origin), back-translation data

Train Data	2-way			3-way			5-way		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
A-origin	0.8055	0.8003	0.8047	0.7481	0.6795	0.7435	0.6875	0.5115	0.6857
B1-backtrans-on-A	0.7796	0.7759	0.7798	0.7185	0.6662	0.7143	0.6444	0.6010	0.6420
B2-caasra-on-A	0.8166	0.8132	0.8167	0.7518	0.7036	0.7495	0.6833	0.5892	0.6859
B3-caasra-on-B1	0.7629	0.7576	0.7625	0.7324	0.6500	0.7185	0.6310	0.5570	0.6317
B4-swap-A-B1	**0.8277**	**0.8225**	**0.8267**	**0.7631**	**0.6934**	**0.7578**	**0.7002**	0.5256	**0.6998**
C1-merge-A-B1	0.8166	0.8124	0.8163	0.7524	0.7102	0.7351	0.6410	**0.6012**	0.6327

Table 3: Our experimental results on 2-way, 3-way, and 5-way of SciEntsBank UA test data.

Method	2-way			3-way			5-way		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
Non-Neural Methods									
CoMeT (Ott et al. 2013)	0.7740	0.7680	0.7730	0.7130	0.6400	0.7070	0.6000	0.4410	0.5980
ETS (Heilman et al. 2013)	0.7760	0.7620	0.7700	0.7200	0.6470	0.7080	0.6430	0.4780	0.6400
SOFTCAR (Jimenez, Becerra, and Cic-Ipn 2013)	0.7240	0.7150	0.7220	0.6590	0.5550	0.6470	0.5440	0.3800	0.5370
Sultan et al. (Sultan, Salazar, and Sumner 2016)	0.7087	0.6768	0.6907	0.6042	0.4439	0.5696	0.4898	0.3298	0.4875
Neural Methods									
Marvaniya et al. (Marvaniya et al. 2018)	0.7700	0.7730	0.7810	0.7000	0.6360	0.7190	0.6025	0.5790	0.6100
Saha et al. (Saha et al. 2018)	0.7926	0.7858	0.7910	0.7185	0.6662	0.7143	0.6444	**0.6010**	0.6420
MDA-ASAS	**0.8277**	**0.8225**	**0.8267**	**0.7631**	**0.6934**	**0.7578**	**0.7002**	0.5256	**0.6998**

Table 4: Comparation of our models with previous state-of-art results on 2-way, 3-way, and 5-way of SciEntsBank UA test data.

Max_sequence_length	Acc	M-F1	W-F1
60	0.8111	0.8074	0.8111
70	0.8111	0.8076	0.8119
80	0.8129	0.8109	0.8136
90	**0.8277**	**0.8225**	**0.8267**
100	0.8189	0.8170	0.8183

Table 5: Ablation study of B4-swap-A-B1 dataset under different maximum sequence lengths settings.

augmentation by original dataset (B1-backtrans-on-A), correct answer as reference answer data augmentation by original dataset (B2-caasra-on-A), correct answer as reference answer data augmentation by twin dataset (B3-caasra-on-B1), swap content data augmentation by original dataset and twin dataset (B4-swap-A-B1), and merge the original dataset and twin dataset (C1-merge-A-B1). The experiments achieve the highest result on the UA test data in the training data of B4-swap-A-B1. Specifically, B4-swap-A-B1 dataset achieves 2 points, 2 points, and 1 point macro-averaged-F1 gains with our original dataset in 2-way, 3-way, and 5-way respectively. For the UA test set, B1-backtrans-on-A dataset obtains 2 point, 1 point, and 1 point weighted-average-F1 gains with our original dataset in 2-way, 3-way, and 5-way respectively. Experiments show that the method of using data augmentation for our training set is effective for short answers scoring.

Ablation Study In order to gain a detailed understanding of the MDA-ASAS, we perform an ablation study on the B4-swap-A-B1 training dataset. The maximum sequence length (max_sequence_length) has a large impact on the fine-tuned BERT model. In this work, we evaluate the performance of the ASAS task with different maximum sequence length settings in 2-way of SciEntsBank UA test data. The results are summarized in Table 5. In data preprocessing, we have analyzed the size of the word bag in the reference answer and the corresponding student answer. And then we sort all the lengths from smallest to largest. Through the analysis of positive distribution in the data, we selected 99.73% of the total number of all lengths and obtained through the index of 4955/4969, which is 80. Therefore, we perform detection experiments on the data around 80. Finally, we conduct a search in the range of 60 to 100 lengths, with each step size of 10.

Table 5 shows that the maximum sequence length is 90 better than the other numbers. Since we start with the reference answer and the student answer as input to our fine-tuned BERT model, when the maximum sequence length is set to 90, it just covers all the information of the sentence. Therefore, the model has a maximum sequence length of 90, which can effectively improve the performance of the ASAS task.

Comparison with State-of-the-Art Models From Table 3, we get the best data augmentation by original data, which is swap content data augmentation by original dataset and twin dataset (B4-swap-A-B1), We regard the B4-swap-A-B1 training dataset as our final MDA-ASAS dataset. Therefore, we compare MDA-ASAS with six state-of-the-art models for ASAS. They include four non-neural models and two neural models. The non-neural models are CoMeT (Ott et al. 2013), ETS (Heilman et al. 2013), SOFTCAR (Jimenez, Becerra, and Cic-Ipn 2013) and Sultan et al. (Sultan, Salazar, and Sumner 2016). CoMeT, ETS, and SOFTCAR are three of the best-performing systems in the SemEval-2013 task. Note that ETS benefits from its underlying domain adapta-

tion. Sultan et al. (Sultan, Salazar, and Sumner 2016) is recent research that proposes a method for short answer scoring in a feature ensemble approach involving text alignment, semantic similarity, question demoting, term weighting, and length ratios.

One of the two neural models is a state-of-the-art model by Marvaniya et al. (Marvaniya et al. 2018). Marvaniya et al. (Marvaniya et al. 2018) utilizes simple lexical baseline features and sophisticated sentence-embedding base features. The other neural model is Saha et al. (Saha et al. 2018). Saha et al. utilize hand-crafted token features along with deep learning embeddings, suggesting that such fusion is helpful for ASAS. Table 4 reports all the results.

We find that the MDA-ASAS yields significantly better results than all compared systems except Saha et al. (Saha et al. 2018) in all the three tasks. We report 4 points and 3 points better macro-averaged-F1 than Saha et al. (Saha et al. 2018) in 2-way and 3-way respectively. For 5-way, both accuracy and weighted-average-F1 metrics are better than all methods, and only macro-averaged-F1 is worse than Saha et al. (Saha et al. 2018). We observe that MDA-ASAS combined with fine-tuned BERT model can improve performance on automatic short answer scoring.

Conclusion and Future Work

To improve the performance of ASAS systems, we propose MDA-ASAS, multiple data augmentation strategies on automatic short answer scoring. Multiple data augmentation strategies consist of back-translation, correct answer as reference answer, and swap content. We have shown that MDA-ASAS combined with fine-tuned BERT model can improve performance on automatic short answer scoring, which exploit the powerful features of external knowledge. In summary, the performance of the ASAS system can benefit from integrating both MDA features and pre-training language model embedding features.

Although our results are specific to the task of ASAS, we believe that the data augmentation strategies of MDA-ASAS can be directly applied to any semantic similarity task that requires capturing external knowledge features. On the other hand, the improvement in our method is sometimes marginal. We only conduct some experiments in the UA test set, we will plan to conduct more experiments in the UQ and UD test set in the future. Continued work on this topic could explore the theoretical underpinning of the data augmentation strategy. We hope that MDA-ASAS's diversity and utility make a compelling case for further thought.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U1811263) and the Guangzhou Key Laboratory of Big Data and Intelligent Education (201905010009).

References

Bergmanis, T.; Kann, K.; Schütze, H.; and Goldwater, S. 2017. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, 31–39.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. 670–680.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dzikovska, M.; Nielsen, R.; Brew, C.; Leacock, C.; Giampiccolo, D.; Bentivogli, L.; Clark, P.; Dagan, I.; and Dang, H. T. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge.

Dzikovska, M. O.; Nielsen, R. D.; and Brew, C. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 200–210. Association for Computational Linguistics.

Hassan, S.; Fahmy, A. A.; and El-Ramly, M. 2018. Automatic short answer scoring based on paragraph embeddings. *International Journal of Advanced Computer Science and Applications* 9(10):397–402.

Heilman; Michael; Madnani; and Nitin. 2013. Ets: Domain adaptation and stacking for short answer scoring.

Jia, R., and Liang, P. 2016. Data recombination for neural semantic parsing. *arXiv preprint arXiv:1606.03622*.

Jimenez, S.; Becerra, C.; and Cic-Ipn, A. G. 2013. Softcardinality: Hierarchical text overlap for student response analysis. In *Joint Conference on Lexical and Computational Semantics*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Ko, T.; Peddinti, V.; Povey, D.; and Khudanpur, S. 2015. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kobayashi, S. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kumar, S.; Chakrabarti, S.; Roy, S.; Kumar, S.; Chakrabarti, S.; and Roy, S. 2017. Earth mover's distance pooling over siamese lstms for automatic short answer grading. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2046–2052.

Liu, J.; Xu, Y.; and Zhao, L. 2019. Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.

Marvaniya, S.; Saha, S.; Dhamecha, T. I.; Foltz, P.; Sindhgatta, R.; and Sengupta, B. 2018. Creating scoring rubric

from representative student answers for improved short answer grading. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 993–1002. ACM.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the Acm* 38(11):39–41.

Mohler, M.; Bunescu, R.; and Mihalcea, R. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Meeting of the Association for Computational Linguistics: Human Language Technologies*, 752–762.

Mueller, J., and Thyagarajan, A. 2016. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2786–2792.

Ott, N.; Ziai, R.; Hahn, M.; and Meurers, D. 2013. Comet: Integrating different levels of linguistic modeling for meaning assessment. In *International Workshop on Semantic Evaluation*, 608–616.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Riordan, B.; Horbach, A.; Cahill, A.; Zesch, T.; and Lee, C. M. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 159–168.

Saha, S.; Dhamecha, T. I.; Marvaniya, S.; Sindhgatta, R.; and Sengupta, B. 2018. Sentence level or token level features for automatic short answer grading?: Use both. In *International Conference on Artificial Intelligence in Education*, 503–517. Springer.

Saha, S.; Dhamecha, T. I.; Marvaniya, S.; Foltz, P.; Sindhgatta, R.; and Sengupta, B. 2019. Joint multi-domain learning for automatic short answer grading. *arXiv preprint arXiv:1902.09183*.

Siddiqi, R.; Harrison, C. J.; and Siddiqi, R. 2010. Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies* 3(3):237–249.

Silfverberg, M.; Wiemerslage, A.; Liu, L.; and Mao, L. J. 2017. Data augmentation for morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection* 90–99.

Sultan, M. A.; Salazar, C.; and Sumner, T. 2016. Fast and easy short answer grading with high accuracy. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1070–1075.

Wang, W. Y., and Yang, D. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2557–2563.

Wei, J. W., and Zou, K. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Wu, X.; Lv, S.; Zang, L.; Han, J.; and Hu, S. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, 84–95. Springer.

Xia, Y.; Qin, T.; Chen, W.; Bian, J.; Yu, N.; and Liu, T.-Y. 2017. Dual supervised learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3789–3798. JMLR. org.

Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, 649–657.

Zhao, S.; Zhang, Y.; Xiong, X.; Botelho, A.; and Heffernan, N. 2017. A memory-augmented neural model for automated grading. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 189–192. ACM.