Pre-Training BERT on Domain Resources for Short Answer Grading

Chul Sung[*]
IBM Watson Education, USA
daniel.c.sung@gmail.com

Tejas Indulal Dhamecha
IBM Research, India
tidhamecha@in.ibm.com

Swarnadeep Saha[†]
UNC Chapel Hill, USA
swarna@cs.unc.edu

Tengfei Ma
IBM Research AI, USA
tengfei.ma1@ibm.com

Vinay Reddy
IBM Watson Education, USA
vinay.kasireddy@ibm.com

Rishi Arora
IBM GBS, India
arishi@in.ibm.com

Abstract

Pre-trained BERT contextualized representations have achieved state-of-the-art results on multiple downstream NLP tasks by fine-tuning with task-specific data. While there has been a lot of focus on task-specific fine-tuning, there has been limited work on improving the pre-trained representations. In this paper, we explore ways of improving the pre-trained contextual representations for the task of automatic short answer grading, a critical component of intelligent tutoring systems. We show that the pre-trained BERT model can be improved by augmenting data from the domain-specific resources like textbooks. We also present a new approach to use labeled short answering grading data for further enhancement of the language model. Empirical evaluation on multi-domain datasets shows that task-specific fine-tuning on the enhanced pre-trained language model achieves superior performance for short answer grading.

1 Introduction

Intelligent tutoring system (ITS) is one of the tools to facilitate personalized learning. Automatic short answer grading (ASAG) is an important component of a dialog-based tutoring (DBT) system; especially, to enable Socratic tutoring. Automatic grading is the task of evaluating the correctness of a student answer for a specific question by comparing it to a reference answer. In short answer grading, the reference answers are typically one or two sentences long and close ended. A broad range of approaches from simple bag-of-words to transfer learning and deep neural networks have been explored to address the short answer grading problem (Mueller and Thyagarajan, 2016; Saha et al., 2018; Marvaniya et al., 2018).

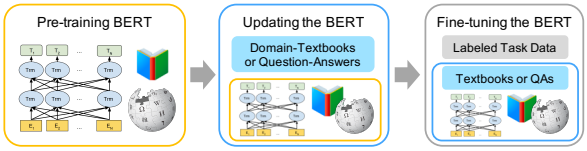

Figure 1: In this work, we propose to update the pre-trained BERT language model by utilizing the textbook or question-answer data to improve short answer grading.

For a variety of natural language processing (NLP) tasks, state-of-the-art results have been reported with pre-trained deep language models, such as BERT (Devlin et al., 2018), GPT (Radford et al., 2018), and ULMFit (Howard and Ruder, 2018). In these approaches, pre-trained language models are utilized for down-stream NLP tasks by means of task-specific *fine-tuning*.

A typical DBT, and therefore a short answer grading system, is often used across various subjects or domains (e.g., Science, Sociology, and Psychology). A pre-trained language model, such as that of BERT, is typically trained on generic English language corpus. Thus, there may be a scope for updating and improving the pre-trained language model on available textual resources for the domains of short answer grading. Textbooks used to create questions and reference answers are mostly available in digital form. Moreover, the labeled data for short answer grading have question and reference answer pairs available but contextual information between the pairs and even just questions are easily ignored. Thus, in this research, we aim to explore and evaluate various methods of updating the pre-trained BERT language model (LM)[*] on such domain-specific

[*]Corresponding author.

[†]The work was done when the author was an employee at IBM Research, India.

[*]We prefer the phrase *to update LM* to imply further training or modifying the pre-trained LM. In some literature, the term *LM fine-tuning* is also used to refer to the update of pre-trained language model. However, we reserve the term *fine-tuning* for task-specific classifier training in sync with BERT terminology.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6071–6075,
Hong Kong, China, November 3–7, 2019.

available resources in the context of ASAG to answer the following research questions.

RQ1 Is updating pre-trained BERT LM helpful in improving short answer grading performance?

RQ2 What is the effect of unsupervised domain corpora (i.e. domain textbooks) in updating LM for short answer grading?

RQ3 How generalizable are the pre-trained and the updated BERT models to unseen domains, in case, where textbooks are not available?

RQ4 How can labeled Question-Answer data be exploited to update LM for short answer grading, in addition to fine-tuning?

Our evaluations are performed on four subjects (Physiology, American Government, and two on Psychology). In addition to the empirical analysis, we also propose a novel approach to effectively utilize the question-answer data as part of the pre-trained model update.

2 Related Work

The problem of short answer grading has attracted significant attention of the researchers over the years. Various approaches, starting from traditional hand-crafted features (Mohler et al., 2011; Sultan et al., 2016) to more recent deep learning models (Riordan et al., 2017; Kumar et al., 2017) and their combination (Saha et al., 2018) have been explored. However, similar to most downstream NLP tasks, ASAG also suffers from the overhead of task-specific architectures and thus scalability across different subjects has proven to be hard.

In a step towards alleviating this overhead, the NLP community has recently proposed multiple generic pre-trained language models which can be transferred seamlessly and fine-tuned for any end task. Universal Language Model Fine-Tuning (ULMFiT) (Howard and Ruder, 2018) method is one of the first such initiatives to illustrate the effectiveness of language model fine-tuning. Embeddings from Language Models, commonly referred to as ELMo (Peters et al., 2018) also learns deep contextualized word representations using the internal states of a deep bidirectional language model. Finally, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) improves on all previous pre-training techniques by training a deep language model that jointly conditions on both the left and right context simultaneously through all the layers. BERT's effectiveness has been widespread and Sung *et al.* (Sung et al., 2019) have shown that fine-tuning BERT for ASAG also outperforms all existing techniques.

While BERT has been fine-tuned to achieve state-of-the-art results on a large number of tasks, the idea of further pre-training of the language model to incorporate more domain knowledge has been explored less. BioBERT (Lee et al., 2019) for biomedical tasks and SciBERT (Beltagy et al., 2019) for science domains have shown the effectiveness of language model pre-training for tasks in specific domains. Motivated by these prior works, we propose two pre-training techniques for BERT in the context of short answer grading that improve over the generic BERT.

3 Method

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a deep bidirectional pre-trained language model that is fine-tuned for downstream NLP tasks. As illustrated in Figure 1, we describe the usage of BERT in two parts: firstly the proposed enhancements in the pre-training step and secondly, the fine-tuning step for short answer grading.

3.1 Pre-Training BERT

We start with the pre-trained BERT language model, trained using the English Wikipedia and BooksCorpus and propose two methods to further improve it.

3.1.1 Usage of Textbooks

Our first approach relies on the usage of textbooks from specific domains of short answer grading. Specifically, we collect textbooks corresponding to the domains and chunk them into paragraphs and feed each paragraph as a document for pre-training. Since our task at hand is short answer grading, we assume that the answers to questions do not overlap between paragraphs and thus we treat each paragraph as an independent document.

To validate our assumption, we randomly sampled 60 question-answer pairs from Physiology domain, and manually examined these whether the answer is contained in the same paragraph of the question. We observe that for about 90% of these

samples, that is indeed the case. Further details on the textbooks and their pre-processing steps for data preparation is provided in the experiments section.

3.1.2 Usage of Question-Answer Pairs

Our second approach leverages the labeled (question, reference answer, student answer) data triples as unsupervised data for pre-training the language model. We consider the triples with correct labels only and create pairs of the following form for each of the correct student answers.

Question-Reference Answer pair
What does the telencephalon contain? The telencephalon contains most of the cerebral cortex.
Question-Correct Answer pair
What does the telencephalon contain? It contains the cerebral cortex, limbic system, and basal ganglia.

Each of these pairs is fed as documents to the BERT architecture. Since one of the training objectives for BERT is next sentence prediction, the answer in a (question, answer) pair provides an immediate context to the question. Incorrect and partially correct student answers, apart from being factually incomplete, are often grammatically incorrect and thus may harm the language model learning. Hence, we ignore those triples for pre-training.

3.2 Fine-tuning for ASAG

The BERT fine-tuning step using labeled short answer grading data proceeds similar to any sentence pair classification task. The (reference answer, student answer) pair is converted to a single sequence of tokens by using a separator token `[SEP]` between the pair and a classification token `[CLS]` at the beginning. The input pair's representation, as obtained from the embedding of the `[CLS]` token, is then fed into a dense layer, which along with the language model is updated during fine-tuning.

4 Experiments

The experiments section is organized as follows. We start by providing a brief description of the dataset, followed by a discussion about the implementation details. In the concluding two subsections, we analyze the effect of the two proposed pre-training methodologies for short answer grading.

Textbook	Train			Test		
	Correct	Incorrect	Partial	Correct	Incorrect	Partial
Phy	9,676	1,197	2,524	4,784	593	1,321
Gov	9,405	3,483	2,433	4,784	1,699	1,177
Psy-I	8,151	736	1,700	4,144	354	795
Psy-II	8,324	755	2,588	4,146	370	1,317

Table 1: Class-wise student answer distribution in train and test sets for the four domains used in experiments.

Corpus	# of words	Domain
English Wiki	2.5B	General
BooksCorpus	0.8B	General
Textbooks	1.1M	Phy + Gov
QAs	0.6M	Phy + Gov
	1.3M	Phy + Gov + Psy-I,II

Table 2: Size and domain of pre-training text corpora.

4.1 Dataset

We show results on a proprietary large-scale industry dataset consisting of three domains - (1) Physiology of Behavior (Phy), (2) American Government (Gov), and (3) Psychology – Human Development (Psy-I) and Abnormal Psychology (Psy-II). Given a question, reference answer and student answer, we address a 3-way classification task into `correct`, `incorrect` and `partially correct` classes. Table 1 shows the train-test splits for each of the domains.

We obtain three textbooks for each of the Physiology and American Government domains. These include our own textbooks and additional ones downloaded from Lumen Learning[†] and Gutenberg[‡] websites. We do not use any textbooks for the Psychology domain, to show the effect of BERT pre-training on out-of-domain data. We combine the data from all the textbooks for further pre-training of BERT. Table 2 summarizes the sizes of the textbook corpora (1.1M words) and question-answer corpora (1.3M words) used in our experiments. Note that, the original BERT model is learned with about 3.3B words corpora.

4.2 Implementation Details

We leverage the TensorFlow implementation of BERT-Base[§] for all our experiments. It is further pre-trained for 240K, 150K, and 240K epochs for BERT+Textbook$_{\text{Phy+Gov}}$, BERT+QA$_{\text{Phy+Gov}}$, and BERT+QA$_{\text{Phy+Gov+Psy-I,II}}$ respectively using the same hyperparameters until the accuracy of the two pre-training objectives converges to 100%.

[†] *https://lumenlearning.com/*
[‡] *https://www.gutenberg.org/*
[§] *https://github.com/google-research/bert*

	BERT	(+Textbook)	(+QA)	(+Textbook+QA)
Phy	81.35	82.28	83.95	83.64
Gov	74.55	76.25	75.43	77.28
Psy-I	78.69	78.05	77.71	77.89
Psy-II	80.06	79.10	79.46	80.74

Table 3: Effect of domain data. Textbook and Question-Answer data from only Phy and Gov domains are used in the pre-trained model update; making Psy-I and Psy-II unseen domains.

Note that the corpus size for Phy+Gov is smaller, leading to faster convergence. Once the pre-training is done, we fine-tune the model with the short answer grading labeled data for 3 epochs using a learning rate of 3e-5. All results are reported in terms of macro-averaged F1.

4.3 Effect of Domains Textbook Data

In this set of experiments, we aim to understand the effect of updating the pre-trained BERT LM on textbook data. We take the textbooks from only two domains (Physiology and American Government), for designing a scenario where textbooks are not available. Such a scenario helps us understand the generalizability of BERT (for which all the domains are unseen), and BERT+Textbook data (for which Psy-I and Psy-II are unseen domains). We combine the data from both the domains and pre-train a single BERT model as pre-training per-domain models is computationally expensive and almost impossible to scale.

Table 3 shows the results. The pre-trained BERT model, as is, performs fairly well (74-81% M-F1). The LM updated with textbook data (BERT+Textbook), improves performance on the domains included in additional pre-training (Phy and Gov). However, we suspect that the updated model becomes more specialized towards seen domains, which leads to performance degradation on the unseen domain of Psychology.

- RQ1 and RQ2 can be answered as *LM update using domain textbook data positively affects the short answer grading performance on the corresponding domains.*
- RQ3 can be answered as *the updated BERT LM model does not appear to generalize well on unseen domains, as the evidence suggests that LM becomes more domain-specific.*

4.4 Effect of Question-Answer Data

In this set of experiments, we aim to understand effectiveness of the Question-Answer (QA) data to update the LM. As explained earlier, for each reference answer and correct answer, question-answer pairs are created and utilized as documents.

In Table 3, a combined QA dataset from Psy and Gov subjects is used for running additional steps of LM pre-training. This again simulates the unseen domain scenario for Psy-I and Psy-II in pre-training update. It can be observed that the proposed approach utilizing QA data (BERT+QA) improves the performance consistently in both the seen subjects of Phy and Gov. Akin to textbook data experiments, the performance is degrading for unseen domains as the model becomes more specialized. Interestingly enough, the updated LM on both strategies (BERT+Textbook+QA) also positively impacts on in-domain performance.

Additionally, another set of results is obtained by utilizing the QA dataset of *all four* domains; simulating an all seen-domain scenario. Table 4 reports the corresponding results showing that the QA data helps improve the model for all the domains, consistently.

	BERT	BERT+QA
Phy	81.35	**83.36**
Gov	74.55	**76.25**
Psy-I	78.69	**78.73**
Psy-II	80.06	**81.41**

Table 4: Effect of Question-Answer data in seen-domains scenario. All four domains are included in the Question-Answer data.

These observations suggest that the proposed approach re-purposes the `Next Sentence Prediction` task to effectively encode the latent features of ASAG. Further, RQ1 can be answered affirmatively; and RQ4 can be answered as *it is advisable to use the QA corpora for LM model update in addition to fine-tuning.*

5 Conclusion

In this paper, we proposed two ways to update the pre-trained BERT language model for the short answer grading task. We illustrate utilization of unstructured textbook data and labeled question answer data for the model update. We show that by adding a step of updating BERT using these domain-related resources, we can achieve better results than directly fine-tuning pre-trained BERT on the end task. We also observed that the updated model becomes more specialized towards the corresponding domains, adversely affecting the per-

formance on unseen domains.

We limited the scope of this paper to the task of automatic short answer grading only. However, our findings of the sensitivity of domain-specific BERT models appear generic. We believe that any multi-domain text classification task should exhibit similar behavior. We also note that our strategies for improving the BERT model should be directly applicable to other QA tasks as well. Future directions include trying our method on different tasks and different relevant data.

References

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Sachin Kumar, Soumen Chakrabarti, and Shourya Roy. 2017. Earth moverâĂŹs distance pooling over siamese lstms for automatic short answer grading. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2046–2052.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Smit Marvaniya, Swarnadeep Saha, Tejas I Dhamecha, Peter Foltz, Renuka Sindhgatta, and Bikram Sengupta. 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*, pages 993–1002. ACM.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 752–762. Association for Computational Linguistics.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, volume 16, pages 2786–2792.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168.

Swarnadeep Saha, Tejas I Dhamecha, Smit Marvaniya, Renuka Sindhgatta, and Bikram Sengupta. 2018. Sentence level or token level features for automatic short answer grading?: Use both. In *International Conference on Artificial Intelligence in Education*, pages 503–517. Springer.

Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075.

Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. In *Artificial Intelligence in Education*, volume 2, pages 469–481. Springer International Publishing.