Towards Improving Open Student Answer Assessment using Pretrained Transformers

Nisrine Ait Khayi, Vasile Rus, Lasang Tamang
Institute for Intelligent Systems/The University of Memphis, Memphis, TN
{ntkhynyn, vrus, ljtamang}@memphis.edu

Abstract

The transfer learning pretraining-finetuning paradigm has revolutionized the natural language processing field yielding state-of the art results in several subfields such as text classification and question answering. However, little work has been done investigating pretrained language models for the open student answer assessment task. In this paper, we fine tune pretrained T5, BERT, RoBERTa, DistilBERT, ALBERT and XLNet models on the DT-Grade dataset which contains freely generated (or open) student answers together with judgment of their correctness. The experimental results demonstrated the effectiveness of these models based on the transfer learning pretraining-finetuning paradigm for open student answer assessment. An improvement of 8%-15% in accuracy was obtained over previous methods. Particularly, a T5 based method led to state-of-the-art results with an accuracy and F1 score of 0.88.

Introduction

Assessment is a key task in education in general and in Intelligent Tutoring Systems (ITSs; Rus et al. 2013) in particular because fully adaptive instruction presupposes accurate assessment (Chi et al. 2001). This assessment provides an estimate of the mastery level of the learner with respect to a target topic. Based on students' mastery level and other important learning-relevant characteristics such as affect and motivation, ITSs trigger appropriate micro-adaptation actions, e.g., in the form of hints and feedback for a particular instructional task, as well as macro-adaptation decisions, e.g., selecting the most appropriate instructional tasks for the student to work on next.

The widely adopted and scalable approach to assessing such open-ended student responses is the so-called semantic (textual) similarity approach. Accordingly, a score, usually normalized, is computed between a target student answer and an expert-provided reference answer (Banjade et al., 2016). If the student answer has a high semantic similarity score to the reference answer, we infer that the student answer has the same correctness value as the reference answer. A low semantic similarity score implies the student response is incorrect. Sometimes, the reference answer provided by experts are common misconceptions. When student responses have a high similarity score to such expert-provided misconceptions then the responses are deemed incorrect, i.e., the student has a major misconception that must be addressed through appropriate feedback and instructional tasks.

Assessing freely generated student' responses in dialog-based systems is challenging as students can express the same idea in different ways owing to different individual style preferences and varied individual characteristics such as cognitive abilities and knowledge. Table 1 shows four answers, articulated by four different college students, to a question asked by a state-of-the-art conversational ITS. It should be noted that all four student answers shown in Table 1 are correct answers to the tutor question. As can be seen from the table, some students write full sentences (student answer A4), some others write very short answers (A3), and yet other students write elaborate answers that include additional concepts relative to the reference answer (A1).

Table 1. Examples of student generated short answers

Problem description: While speeding up, a large truck pushes a small compact car. **Tutor question**: How do the magnitudes of forces they exert on each other compare? **Reference answer:** The forces from the truck and car are equal and opposite. **Student answers:** A1. *The magnitudes of the forces are equal and opposite to each other due to Newton's third law of motion.* A2. *they are equal and opposite in direction* A3. *equal and opposite* A4. *the truck applies an equal and opposite force to the*

This diversity in the level of completeness with respect to the benchmark or reference answer, i.e., the level of information in the reference answer that a student explicitly articulates, makes assessment of such responses using a semantic similarity approach challenging. When information is implied, the use of context and other knowledge sources is needed to recover the implied information.

In the recent past, researchers have made significant progress in solving the open student answer assessment task while accounting for context and other knowledge sources during tutorial dialogues.

using deep learning methods (Khayi et al.,2019, Khayi et al.,2020, Maharjan et al.,2018, Gong et al.,2019). The success of deep learning methods depends on the availability of large amount of high-quality labeled data. In many cases, including ours, the size of available data is small. An option to alleviate this limitation is using transfer learning models, the main focus of our work presented here, as a way to incorporate knowledge from other sources.

The main idea behind transfer learning is to pretrain a model on large amounts of unlabeled data in order to obtain a powerful language model which can then be specialized for solving specific NLP downstream tasks by adding new layers and training them on the target data. These pretrained language models have been used recently to obtain state-of-the-art results in many NLP tasks (Devlin et al., 2019; Yang et al., 2019; Dong et al., 2019; Liu et al., 2019; Lan et al., 2019). Motivated by these successes , we explore the potential of finetuning several pretrained transformers on the student answers assessment downstream task. We experimented with such pretrained transformers on the DT-Grade dataset (Banjade et al. 2016) which contains 900 instances categorized in four classes: correct (367 instances), incorrect (238 instances), correct but incomplete (210 instances), and contradictory (84 instances). To overcome the problem of class size imbalance in the dataset and given its relatively small size, we considered a binary classification where all instances in the incorrect, correct but incomplete, and contradictory categories are deemed as incorrect.

Related Work

The student answer assessment task has attracted broad attention recently. Several researchers have explored the potential of pretrained transformers as they led to state of-the-art results in numerous NLP tasks. For example, Camus and colleagues (2020) experimented with fine tuning multiple pretrained transformers for the automatic short answer grading (ASAG) downstream task, which is related to our task, on the SemEval-2013 dataset. They also investigated the impact of transfer learning from the Multi-genre Natural Language Inference (MNLI) dataset to SemEval-2013 dataset on performance and generalization. The experimental results showed a significant gain of 15% improvement in performance score. The results also showed that distilled versions of the pretrained models with reduced parameters led to a slight decrease in the performance score but still feasible for the ASAG task.

Working on the same task using our DT-Grade dataset, Candor (2020) finetuned BERT on the ASAG downstream task. The model has been evaluated using Cohen's Kappa as a measure of inter-rater reliability between the automated system and the human rater. The Experimental results showed that pretrained models such as BERT can help achieve more consistent human ratings. In their research efforts to improve the performance results of the ASAG task, Sung and colleagues (2019) proposed new ways to enhance the performance of BERT. To this end, they proposed to pretrain BERT on domain specific data such as textbooks and use labeled automatic short answer grading data to enhance the language model. Then, they finetuned the pretrained BERT model on the downstream task by considering two inputs: the student answer and the reference answer. The experimental results showed that fine tuning BERT using the enhanced pretrained model achieves superior performance on the ASAG downstream task.

In this paper, we explore for the first time the potential of pretrained transformers such as T5 and XLNET models and others for the open student answer assessment task.

Methods

BERT (Devlin et al., 2019): pretrains deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts in all layers. The model is trained on the Book Corpus (Zhu et al.,2015) and English Wikipedia. The pair of sentences (student answer/A and reference answer/B) is packed into a single sequence and separated with a special token ([SEP]) and a classification token [CLS] at the beginning. An additional learned embedding is added to every token indicating whether it belongs to sentence A or sentence B. The resulted embedding H of the [CLS] token is then fed into a SoftMax layer that predicts the probability of classification label c.

T5 (Raffel et al.,2019): transforms all NLP tasks into a text-to-text format where the inputs and outputs are text strings. The model was pre-trained on the Colossal Clean Crawled Corpus (C4). The pre-training objective of T5 is similar to BERT's with a small modification which is utilizing a Masked Language Model that masks 15% of the input tokens and replaces them with multiple mask key words instead of a unique one as in the case of BERT. Then, the model is trained to recover the masked tokens. T5 model's architecture is based on both the encoder and decoder of the transformer (Vaswani et al., 2017).

RoBERTa (Liu et al.,2019): retrains BERT with an improved training methodology which involves 1,000% more data and computation power. RoBERTa has a different encoding mechanism from BERT. That is, the student answer and reference answer are separated with two [SEP] tokens and a [CLS] token is added at the beginning.

XLNet (Yang et al.,2019): uses a permutation-based language modeling objective to capture bidirectional contexts while retaining the benefits of autoregressive (AR) models. Permutation language models are trained to predict one token given preceding context in some random order. Unlike the other previous pretrained models, the architecture of XLNet is based on the XL-Transformer (Dai et al.,2019). Similar to the finetuning used for BERT, we concatenate the student answer and reference answers separated with a [SEP] token. The [CLS] is added at the end.

DistilBERT (Sanh et al., 2019): uses a technique called distillation which approximates the large model of BERT with a smaller one. DistilBERT is distilled on very large batches by leveraging gradient accumulation using dynamic masking and without the next sentence prediction objective. The experiments have demonstrated a high impact of this reduction on computation efficiency. The encoding mechanism is similar to BERT.

ALBERT (Lan et al., 2019): has the same architecture as BERT. It implements two design changes that yields a model with 12M parameters and 89% parameter reduction compared to the BERT model. This results in an efficiency improvement versus a minor performance degradation. The encoding mechanism of the student answer and reference answer is similar to the one we presented earlier for BERT.

Experiments and Results

We conducted experiments with the above-described methods using the DT-Grade dataset (Banjade et al., 2016) that was created by extracting student responses from logged tutorials interactions between 36 junior level college students and a state of the art conversational ITS. Students were asked to solve 9 conceptual Physics problems and then the ITSs offered help as needed through tutorial conversation. The dataset consists of 900 instances consisting of: (i) the Physics problem description, (ii) the prior tutor question, (iii) the student answer to the prior tutor question and (iv) the reference answer.

We performed all our experiments using a Tesla K80 GPU and a total of 12 GB of RAM. All the models were implemented using the HuggingFace's library (Wolf at al., 2019). We used the base versions of the pretrained transformers that are trained for 4 epochs. The Adam optimizer (Kingma et al., 2014) with a learning rate of 3e-5 was used and the gradients were clipped to 1.0 to prevent exploding gradients. We evaluated our models using the Sparse Categorical Cross-entropy loss and the Sparse Categorical accuracy. About 80% of data was used for training and 20% for testing. Each experiment was repeated 100 times with increased random seeds in an attempt to increase the models' performance (Dodge et al., 2020). We report the best performance results across the 500 conducted experiments.

Table 2 shows performance results of finetuning the pretrained transformers on the DT-Grade dataset. As shown in the table, all the pretrained models outperform the previous methods with a significant margin. The T5 transformer has achieved the highest performance with an accuracy of 0.88 and an F1 score of 0.88. The results also showed that the distillation of BERT parameters is feasible for the student answers assessment task. ALBERT and DistilBERT have performed less than other pretrained transformers with an accuracy of 0.80 and an F1 score of 0.80. But still, it is a very good result in comparison with previous methods such as Bi-GRU-Capsnet (Ait Khayi et al., 2019), an attention-based transformer (Ait Khayi et al.,2020), and a Graph Convolutional Network (Ait Khayi et al.,2020). Another observation can be made from the obtained results is that BERT outperforms XLNET which works better for longer sequences, which is not the case of our student and reference answer which are relatively short.

Model	Accuracy	F1
BERT	0.86(+0.13)	0.86
RoBERTa	0.87(+0.14)	0.87
T5	**0.88(+0.15)**	**0.88**
XLNET	0.84(+0.11)	0.84
ALBERT	0.80(+0.7)	0.80
DistilBERT	0.80(+0.7)	0.80
Graph Convolutional Network	0.73	0.73
Bi-GRU-Capsnet+ELMo	0.72	0.70
Transformer+ELMo	0.71	0.70
LSTM+Glove	0.60	0.60

Table 2. Performance results of the pretrained models.

Overall, the experiments have demonstrated that these pretrained transformers assess correctly the very short answers in comparison with previous methods. For example, RoBERTa handles the assessment of very short student answers concatenated with reference answers with a small number of words (average of 10.5 words) better than the Bi-GRU-Capsnet network.

During the experiments, we investigated whether the learning rate and the sequence length parameters have an impact on the performance results. The experimental results showed that the smaller the learning rate the better the performance results. The value of 3e-5 has led to the best results versus the values of 4e-5 and 5e-5. The results also showed that the longer the length of the input sequence the better the performance results.

Conclusions

Several research studies have demonstrated the effectiveness of the transfer learning pretraining-finetuning paradigm for low resource scenarios in NLP as it is the case for the open student answer assessment task. Motivated by these successes with small datasets, we explored the potential of several pretrained transformers on the student answers assessment downstream task. To this end, we finetuned T5, XLNET, BERT, DistilBERT, ALBERT and RoBERTa transformers on the DT-Grade dataset for the first time. The experimental results showed the effectiveness of

these pretrained transformers that surpassed all the previous methods with a significant margin. The T5 transformer has achieved the highest performance with an accuracy of 0.88 and an F1 score of 0.88. This is a new state of the art on the DT-Grade dataset.

In the future, we plan to find better strategies to fine tune and pretrain these transformers on domain related data in order to improve the assessment results.

References

Ait Khayi, N., and Rus, V. 2019. BI-GRU Capsule Networks for Student Answers Assessment. Paper presented at 2019 KDD Workshop on Deep Learning for Education (DL4Ed), Anchorage, Alaska, August 5,2019.

Ait Khayi, N., and Rus, V. 2020. Attention Based Transformer for Student Answers Assessment. *In Proceedings of the Thirty-Third International FLAIRS Conference* (FLAIRS-32), North Miami Beach, Florida, USA, from May 17-20, 2020.

Banjade, R., Maharjan, N., Niraula, N. B., Gautam, D., Samei, B., and Rus, V. 2016. Evaluation Dataset (DT-Grade) and Word Weighting Approach Towards Constructed Short Answers Assessment in Tutorial Dialogue Context. *In Proceedings of the 11th Workshop on* Innovative Use of NLP for Building Educational Applications, San Diego, CA. June 16, 2016. (pp. 182-187).

Camus, L., and Filighera, A. 2020. Investigating Transformers for Automatic Short Answer Grading. *In Proceedings of the International Conference on Artificial Intelligence in Education,* online, from June 6-10,2020. (pp. 43-48) Springer, Cham.

Chi, M., Koedinger, K., Gordon, G., Jordan, P., and Vanlehn, K. 2011. Instructional Factors Analysis: A Cognitive Model For Multiple Instructional Interventions. *In Proceedings of the 4th International Conference on Educational Data Mining,* Eindhoven, Netherlands, from July -8,2011 pp. 61–70

Condor, A. 2020. Exploring Automatic Short Answer Grading as a Tool to Assist in Human Rating. *In Proceedings of the International Conference on Artificial Intelligence in Education*, online, from June 6-10,2020 (pp. 74-79). Springer, Cham.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N.A. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. Computing Research Repository, arXiv: abs/2002.06305.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Hon, H. W. 2019. Unified language model pre-training for natural language understanding and generation. In Advances in Neural Information Processing Systems (pp. 13042-13054).

Gong. T. and Yao. X. 2019. An Attention-based Deep Model for Automatic Short Answer Score. *International Journal of Computer Science and Software Engineering,* 8(6), 127-132.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of NAACL*, Minneapolis, Minnesota, USA, from June 2-7,2019.

Kingma, D. P. and Ba, J. 2014. Adam: A method for stochastic optimization. *Research Repository*, arXiv:1412.6980.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv:1909.11942.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Research Repository, arXiv: abs/1907.11692.

Maharjan, N., Gautam, D. and Rus, V. 2018. Assessing free student answers in tutorial dialogues using LSTM models. *In Proceedings of AIED, the International* Conference on Artificial Intelligence in Education (pp. 193-198). Springer, Cham, London, UK, from June 25-30, 2018.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Research Repository, arXiv: abs/1910.01108.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ...and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Rus, V., D'Mello, S. K., Hu, X. and Graesser, A. C. 2013. Recent advances in intelligent tutoring systems with conversational dialogue, *AI Magazine,* 34(3), 42-54

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., and Arora, R. 2019. Pre-Training BERT on Domain Resources for Short Answer Grading. *In Proceedings of EMNLP-IJCNLP* (pp. 6073-6077), Hong Kong, CHINA, from November 3-7, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. 2017. Attention Is All You Need. *In Proceedings of NIPS*, Long Beach, CA, USA, from December 4-9,2017.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... and Brew, J. 2019. Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., and Le, Q.V.2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Advances in neural information processing systems (pp. 5754-5764).