Investigating Transformers for Automatic Short Answer Grading

Leon Camus(✉) and Anna Filighera

TU Darmstadt, Darmstadt, Germany
camus@algo.informatik.tu-darmstadt.de, anna.filighera@kom.tu-darmstadt.de

Abstract. Recent advancements in the field of deep learning for natural language processing made it possible to use novel deep learning architectures, such as the Transformer, for increasingly complex natural language processing tasks. Combined with novel unsupervised pre-training tasks such as masked language modeling, sentence ordering or next sentence prediction, those natural language processing models became even more accurate. In this work, we experiment with fine-tuning different pre-trained Transformer based architectures. We train the newest and most powerful, according to the glue benchmark, transformers on the SemEval-2013 dataset. We also explore the impact of transfer learning a model fine-tuned on the MNLI dataset to the SemEval-2013 dataset on generalization and performance. We report up to 13% absolute improvement in macro-average-F1 over state-of-the-art results. We show that models trained with knowledge distillation are feasible for use in short answer grading. Furthermore, we compare multilingual models on a machine-translated version of the SemEval-2013 dataset.

Keywords: Self-attention · Transfer learning · Short answer grading

1 Introduction

Online tutoring platforms enable students to learn individually and independently. To provide the users with individual feedback on their answers, the answers have to be graded. In large tutoring platforms, there are an abundant number of domains and questions. This makes building a general system for short answer grading challenging, since domain-related knowledge is frequently needed to evaluate an answer. Additionally, the increasing accuracy of short answer grading systems makes it feasible to employ them in examinations. In this scenario it is desirable to achieve the maximum possible accuracy, with a relatively high computational budget, while in case of tutoring a less computational intensive model is desirable to keep costs down and increase responsiveness. In this work, we experiment with fine-tuning the most common transformer models and explore the following questions:

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 43–48, 2020.
https://doi.org/10.1007/978-3-030-52240-7_8

Does the size of the Transformer matter for short answer grading? How well do multilingual Transformers perform? How well do multilingual Transformers generalize to another language? Are there better pre-training tasks for short answer grading? Does knowledge distillation work for short answer grading?

The field of short answer grading can mainly be categorized into two classes of approaches. The first ones represent the traditional approaches, based on handcrafted features [14,15] and the second ones are deep learning based approaches [1,8,13,16,18,21]. One of the core constraints of short answer grading remained the limited availability of labeled domain-relevant training data. This issue was mitigated by transfer learning from models pre-trained using unsupervised pre-training tasks, as shown by Sung et al. [21] outperforming previous approaches by about twelve percent. In this study, we aim to extend upon the insights provided by Sung et al. [21].

2 Experiments

We evaluate our proposed approach on the SemEval-2013 [5] dataset. The dataset consists of questions, reference answers, student answers and three-way labels, represenenting the CORRECT, INCORRECT and CONTRADICTORY class. We translate it with the winning method from Wmt19 [2]. For further information see Sung et al. [21]. We also perform transfer learning from a model previously fine-tuned on the MNLI [22] dataset.[1]

For training and later comparison we utilize a variety of models, including BERT [4], RoBERTa [11], AlBERT [10], XLM [9] and XLMRoBERTa [3]. We also include distilled models of BERT and RoBERTa in the study [19]. Furthermore we include a RoBERTa based model previously fine-tuned on the MNLI dataset.

For fine tuning we add a classification layer on top of every model. We use the AdamW [12] optimizer, with a learning rate of 2e−5 and a linear learning rate schedule with warm up. For large transformers we extend the number of epochs to 24, but we also observe notable results with 12 epochs or less. We train using a single NVIDIA 2080ti GPU (11 GB) with a batch size of 16, utilizing gradient accumulation. Larger batches did not seem to improve the results. To fit large transformers into the GPU memory we use a combination of gradient accumulation and mixed precision with 16 bit floating point numbers, provided by NVIDIAs apex library[2]. We implement our experiments using huggingfaces transformer library [23]. We will release our training code on GitHub[3]. To ensure comparability, all of the presented models where trained with the same code, setup and hyper parameters (Table 1).

3 Results and Analysis

Does the size of the Transformer matter for short answer grading? Large models demonstrate a significant improvement compared to Base models.

[1] https://www.nyu.edu/projects/bowman/multinli/.
[2] https://github.com/NVIDIA/apex.
[3] https://github.com/28Smiles/SAS-AIED2020.

Table 1. Results on the SciEntsBank Dataset of SemEval 2013. Accuracy (Acc), macro-average-F1 (M-F1), and weighted-average-F1 (W-F1) are reported in percentage.

	Languages Trained	English									German								
		Unseen answer			Unseen question			Unseen domain			Unseen answer			Unseen question			Unseen domain		
		Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
Baseline [5]	en	55.6	40.5	52.3	54.0	39.0	52.0	57.7	41.6	55.4	-	-	-	-	-	-	-	-	-
ETS [6]	en	72.0	64.7	70.8	58.3	39.3	53.7	54.3	33.3	46.1	-	-	-	-	-	-	-	-	-
SOFTCAR [7]	en	65.9	55.5	64.7	65.2	46.9	63.4	63.7	48.6	62.0	-	-	-	-	-	-	-	-	-
MEAD [17]	en	-	42.9	55.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Graph [17]	en	-	43.8	56.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sultan et al. [20]	en	60.4	44.4	57.0	64.3	45.5	61.5	62.7	45.2	60.3	-	-	-	-	-	-	-	-	-
Saha et al. [18]	en	71.8	66.6	71.4	61.4	49.1	62.8	63.2	47.9	61.2	-	-	-	-	-	-	-	-	-
Marvaniya et al. [13]	en	-	63.6	71.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sung et al. [21]	en	75.9	72.0	75.8	65.3	57.5	64.8	63.8	57.9	63.4	-	-	-	-	-	-	-	-	-
BERT$_{\text{DISTILL}}$	en	69.2	67.2	69.2	56.6	54.7	56.6	61.4	49.7	61.4	38.8	26.2	38.8	33.7	21.5	33.7	51.7	29.5	51.7
BERT$_{\text{BASE}}$	en	72.8	70.6	72.8	57.3	56.0	57.3	63.4	54.6	63.4	45.0	37.0	45.0	40.5	32.9	40.5	59.0	41.7	59.0
BERT$_{\text{LARGE}}$	en	75.8	75.0	75.8	63.4	62.4	63.4	67.7	62.8	67.7	50.2	40.5	50.2	43.7	35.5	43.7	57.7	42.2	57.7
RoBERTa$_{\text{DISTILL}}$	en	74.8	73.2	74.8	56.9	55.2	56.9	65.1	55.6	65.1	48.0	40.4	48.0	44.1	36.7	44.1	58.2	40.2	58.2
RoBERTa$_{\text{BASE}}$	en	74.5	73.2	74.5	63.2	61.7	63.2	65.3	62.5	65.3	47.8	38.1	47.8	43.5	35.1	43.5	60.6	43.2	60.6
RoBERTa$_{\text{LARGE}}$	en	76.7	75.5	76.7	64.1	62.7	64.1	66.8	65.6	66.8	48.8	40.4	48.8	42.7	34.7	42.7	61.5	48.2	61.5
RoBERTa$_{\text{LARGE}}$	de	41.2	19.4	41.2	47.7	21.5	47.7	42.0	19.7	42.0	41.2	19.4	41.2	47.7	21.5	47.7	42.0	19.7	42.0
RoBERTa$_{\text{LARGE}}$	en, de	76.1	74.9	76.1	63.0	61.9	63.0	65.6	63.3	65.6	73.9	72.3	73.9	58.9	57.3	58.9	61.9	56.3	61.9
RoBERTa$_{\text{LARGE,MNLI}}$	en	78.8	78.3	78.8	**66.4**	**65.7**	**66.4**	**71.8**	**70.8**	**71.8**	52.6	49.3	52.6	46.1	42.5	46.1	60.1	51.9	60.1
RoBERTa$_{\text{LARGE,MNLI}}$	de	62.6	59.1	62.6	55.1	51.5	55.1	66.5	66.8	66.5	74.9	74.0	74.9	**60.8**	**59.0**	**60.8**	62.5	57.2	62.5
RoBERTa$_{\text{LARGE,MNLI}}$	en, de	**79.7**	**79.1**	**79.7**	66.3	65.3	66.3	69.4	69.1	69.4	**76.0**	**75.0**	**76.0**	59.6	58.4	59.6	**64.9**	**59.2**	**64.9**
AlBERT$_{\text{BASE}}$	en	72.6	71.4	72.6	57.6	55.2	57.6	60.1	52.3	60.1	37.0	31.5	37.0	30.0	24.9	30.0	40.6	31.8	40.6
AlBERT$_{\text{LARGE}}$	en	71.3	70.1	71.3	58.1	56.8	58.1	65.3	60.7	65.3	45.0	42.1	45.0	38.7	34.9	38.7	58.1	46.0	58.1
XLM$_{\text{MLM-TLM-XNLI}}$	en	72.6	71.2	72.6	57.6	55.5	57.6	56.3	44.8	56.3	48.0	47.4	48.0	43.7	43.3	43.7	47.7	41.5	47.7
XLM$_{\text{MLM-TLM-XNLI}}$	de	57.0	54.8	57.0	43.7	41.9	43.7	56.4	41.2	56.4	68.8	66.5	68.8	54.0	51.6	54.0	55.8	44.8	55.8
XLM$_{\text{MLM-TLM-XNLI}}$	en, de	64.8	62.2	64.8	52.1	49.2	52.1	48.6	35.7	48.6	63.8	61.2	63.8	51.8	49.7	51.8	48.1	33.8	48.1
XLMRoBERTa$_{\text{BASE}}$	en	75.4	73.8	75.4	59.9	57.9	59.9	62.6	54.4	62.6	64.2	60.6	64.2	52.7	48.3	52.7	60.1	49.5	60.1
XLMRoBERTa$_{\text{BASE}}$	de	69.0	67.4	69.0	53.6	51.9	53.6	62.3	51.9	62.3	73.4	71.7	73.4	56.9	55.6	56.9	60.8	49.7	60.8
XLMRoBERTa$_{\text{BASE}}$	en, de	74.1	72.4	74.1	59.1	57.5	59.1	60.1	48.1	60.1	73.1	71.3	73.1	56.1	54.6	56.1	59.4	46.5	59.4

The improvement arises most likely due to the increased capacity of the model, as more parameters allow the model to retain more information of the pre-training data.

How well do multilingual Transformers perform? The *XLM* [9] based models do not perform well in this study. The *RoBERTa* based models (*XLM-RoBERTa*) seem to generalize better than their predecessors. *XLMRoBERTa* performs similarly to the base *RoBERTa* model, falling behind in the unseen questions and unseen domains category. Subsequent investigations could include fine-tuning the large variant on MNLI and SciEntsBank. Due to GPU memory constraints, we were not capable to train the large variant of this model.

How well do multilingual Transformers generalize to another language? The models with multilingual pre-training show stronger generalization across languages than their English counterparts. We are able to observe that the score of the multilingual model increases across languages it was never fine-tuned on, while the monolingual model does not generalize.

Are there better pre-training tasks for short answer grading? Transfer learning a model from MNLI yields a significant improvement over the same version of the model not fine-tuned on MNLI. It improves the models ability to generalise to a separate domain. The models capabilities on the german version of the dataset are also increased, despite the usage of a monolingual model. The reason for this behavior should be further investigated.

Does knowledge distillation work for short answer grading? The usage of models pre-trained with knowledge distillation yields a slightly lower score. However, since the model is 40% smaller, a maximum decrease in performance of about 2% to the previous state of the art may be acceptable for scenarios where computational resources are limited.

4 Conclusion and Future Work

In this paper we demonstrate that large Transformer-based pre-trained models achieve state of the art results in short answer grading. We were able to show that models trained on the MNLI dataset are capable of transferring knowledge to the task of short answer grading. Moreover, we were able to increase a models overall score, by training it on multiple languages. We show that the skills developed by a model trained on MNLI improve generalization across languages. It is also shown, that cross lingual training improves scores on SemEval2013. We show that knowledge distillation allows for good performance, while keeping computational costs low. This is crucial in evaluating answers from many users, like in online tutoring platforms.

Future research should investigate the impact of context on the classification. Including the question or its source may help the model grade answers, which were not considered during the reference answer creation.

Acknowledgements. We would like to thank Prof. Dr. rer. nat. Karsten Weihe, M.Sc. Julian Prommer, the department of didactics and Nena Marie Helfert, for supporting and reviewing this work.

References

1. Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. arXiv preprint arXiv:1606.04289 (2016)
2. Barrault, L., et al.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 1–61. Association for Computational Linguistics, Florence, August 2019. http://www.aclweb.org/anthology/W19-5301
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dzikovska, M.O., et al.: Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. NORTH TEXAS STATE UNIV DENTON, Tech. rep. (2013)
6. Heilman, M., Madnani, N.: Ets: Domain adaptation and stacking for short answer scoring. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 275–279 (2013)
7. Jimenez, S., Becerra, C., Gelbukh, A.: Softcardinality: Hierarchical text overlap for student response analysis. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 280–284 (2013)
8. Kumar, S., Chakrabarti, S., Roy, S.: Earth mover's distance pooling over siamese lstms for automatic short answer grading. In: IJCAI, pp. 2046–2052 (2017)
9. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
11. Liu, Y., et al.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
12. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 (2017)
13. Marvaniya, S., Saha, S., Dhamecha, T.I., Foltz, P., Sindhgatta, R., Sengupta, B.: 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, pp. 993–1002 (2018)
14. Mohler, M., Bunescu, R., Mihalcea, R.: 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, pp. 752–762. Association for Computational Linguistics (2011)

15. Mohler, M., Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 567–575 (2009)
16. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
17. Ramachandran, L., Foltz, P.: Generating reference texts for short answer scoring using graph-based summarization. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 207–212 (2015)
18. Saha, S., Dhamecha, T.I., Marvaniya, S., Sindhgatta, R., Sengupta, B.: Sentence level or token level features for automatic short answer grading?: use both. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 503–517. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93843-1_37
19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: NeurIPS EMC2 Workshop (2019)
20. Sultan, M.A., Salazar, C., Sumner, T.: 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, pp. 1070–1075 (2016)
21. Sung, C., Dhamecha, T.I., Mukhi, N.: Improving short answer grading using transformer-based pre-training. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 469–481. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_39
22. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112–1122. Association for Computational Linguistics (2018). http://aclweb.org/anthology/N18-1101
23. Wolf, T., et al.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv arXiv:1910.03771 (2019)