Pre-training with Scientific Text Improves Educational Question Generation (Student Abstract)

Hamze Muse*, **Sahan Bulathwela*** **and Emine Yilmaz**

Centre for Artificial Intelligence, University College London
Gower Street, London WC1E 6BT, UK
{hamze.muse.20, m.bulathwela, emine.yilmaz}@ucl.ac.uk

Abstract

With the boom of digital educational materials and scalable e-learning systems, the potential for realising AI-assisted personalised learning has skyrocketed. In this landscape, the automatic generation of educational questions will play a key role, enabling scalable self-assessment when a global population is manoeuvring their personalised learning journeys. We develop *EduQG*, a novel educational question generation model built by adapting a large language model. Our initial experiments demonstrate that *EduQG* can produce superior educational questions by pre-training on scientific text.

Introduction

While digital learning resources are created in abundance (Bulathwela et al. 2020), providing related questions to these resources facilitates self-testing, a critical element of self-regulated learning. Also, question-answering enables an intelligent tutor to reliably verify learner skill mastery, making scalable educational question generation essential for democratising education (Bulathwela et al. 2021; Zhang et al. 2021). While existing language models are being used for question generation, their utility to education is only being explored very recently (Wang et al. 2022). In particular, pre-training large language models with educational text to improve question generation is an unexplored area. This work validates if additional training with educational text can improve questions generated in the educational context. We develop an experiment to adapt a large language model to test this and propose *EduQG*, a novel model for educational question generation. Our initial comparisons with a baseline question generation model indicate that this additional training can improve performance.

Related Work

Prior work mainly utilises i) rule-based and ii) neural-based models for question generation (QG), while neural approaches have dominated the state of the art on QG in different applications including intelligent tutoring (Zhang et al. 2021). When it comes to leveraging QG for education, Leaf system (Vachev et al. 2022) is one of the latest proposed methods. Leaf is a cutting-edge question generation system that fine-tunes a large language model for question generation and multiple-choice distracter generation. Due to the recency and relevance of the Leaf system, we use the QG model of Leaf as the baseline model of this study. Like many cutting-edge models, Leaf uses the *SQuAD 1.1* (Rajpurkar et al. 2016), a reading comprehension dataset containing more than 100,000 questions crowd-sourced on a number of Wikipedia articles, to train the QG component of the system. It does so by fine-tuning a pre-trained T5 language model (Raffel et al. 2020).

Although Leaf has been built for educational use cases using the SQuAD dataset, the SQuAD dataset itself contains questions that are aimed at English reading comprehension. Thus, it is not a strong candidate for testing the question generation capability for more rigorous subject domains such as the sciences. On the contrary, SciQ (Welbl, Liu, and Gardner 2017) is a collection of 13,679 crowdsourced scientific exam questions that includes questions regarding physics, chemistry and other sciences. Although small in comparison to SQuAD, the SciQ dataset is a more relevant dataset that can be used to evaluate the educational QG capabilities of a model. Therefore, we use the SciQ dataset to evaluate the question generation models we build in this work.

While large language models capture a lot of information about the world (Raffel et al. 2020), these models need to be pre-trained further in domain-specific datasets to improve their knowledge and fluency in specific domains (e.g. medicine (Xu, Van Durme, and Murray 2021)). In the realm of scientific information, *S2ORC* is a corpus that consists of 81.1 million scholarly publications in English from various academic fields bringing together the largest publicly accessible collection of machine-readable academic literature to date (Lo et al. 2020). To test our hypothesis that pretraining the model with scientific/academic text would improve its educational QG capability, we use the S2ORC dataset.

Our Approach

The primary objective of our study is to validate if further fine-tuning a system on educational data can improve educational QG. The experiment we set up is illustrated in figure 1. The foundational language model to both our training settings is the T5 language model (Raffel et al. 2020). We

*These authors contributed equally.

Model	Predictive Performance					Linguistic Quality		
	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	F1-Score ↑	Perplexity ↓	Diversity ↑	Grammar Errors ↓
Leaf (Baseline)	27.07	20.22	17.17	16.46	30.90	**30.82**	0.735	**0.102**
EduQG (Ours)	**29.19**	**21.69**	**18.03**	**16.76**	**33.18**	34.36	**0.749**	0.122

Table 1: Comparison of predictive performance and linguistic quality between Leaf (baseline) and EduQG (our proposal). The superior performance is indicated in **bold** face.

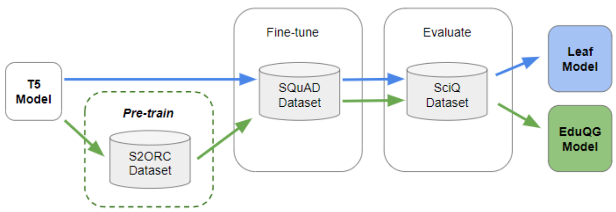

Figure 1: Baseline (blue arrows) and EduQG (green arrows).

first replicate the QG component of the Leaf system (Vachev et al. 2022) by taking the T5 model and fine-tuning it on the SQuAD 1.1 dataset as our baseline QG system (Blue flow in figure 1). As the enhanced proposal, we use the same procedure, except, we fine-tune the T5 model with a down-sampled version of the S2ORC dataset that contains approx. 23.2M scientific abstracts related to Chemistry, Biology and Physics research papers (green dashed box in figure 1).

Evaluation The two settings lead to the baseline (Leaf) and the proposed model (EduQG) that we compare using the SciQ dataset, as it contains exclusively educational questions. To measure the predictive power of the human-generated questions, we use the BLUE score and the F1 score (Rajpurkar et al. 2016). To measure how human-like the generated questions are, we use perplexity, diversity and grammatical error rates. A lower perplexity score indicates better coherence (Wang et al. 2022).

Preliminary Results and Discussion

The results of the model comparison are presented in table 1. The predictive performance results in Table 1 clearly indicate that the *EduQG* model is better at predicting scientific questions based on the context compared to Leaf. This is a strong indication that the additional scientific knowledge the EduGQ model is pre-trained on has an effect on educational QG capability. However, the linguistic quality metrics (shown on the right of the table) do not yield a favourable result although diversity has been improved by our model. We hypothesise that this may be due to the mismatch of language style and vocabulary of a scientific language that is advanced and complex. Therefore, scientific language might not align seamlessly with the reference models used for linguistic quality assessment.

Conclusion

This work introduces EduQG, a foundational step toward further pre-training to improve educational QG. Our initial experiments prove the utility of pre-training an existing language model to improve its performance. The linguistic quality metrics are not as favourable as expected. Deeper analyses are warranted to understand whether the outcomes portray a limitation or a mismatch between the language models which will be addressed in future work using both offline and human studies.

Acknowledgments

This work is funded by the European Commission project ”Humane AI” (grant 820437).

References

Bulathwela, S.; Perez-Ortiz, M.; Yilmaz, E.; and Shawe-Taylor, J. 2020. TrueLearn: A Family of Bayesian Algorithms to Match Lifelong Learners to Open Educational Resources. In *AAAI Conference on Artificial Intelligence*.

Bulathwela, S.; Pérez-Ortiz, M.; Holloway, C.; and Shawe-Taylor, J. 2021. Could AI Democratise Education? Socio-Technical Imaginaries of an EdTech Revolution. In *Proc. of NeurIPS Workshop on ML4D*. arXiv.

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proc. of the Annual Meeting of the ACL*. Online.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J.; et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140): 1–67.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *CoRR*, abs/1606.05250.

Vachev, K.; Hardalov, M.; Karadzhov, G.; Georgiev, G.; Koychev, I.; and Nakov, P. 2022. Leaf: Multiple-choice question generation. In *Proc. of the European Conf. on Information Retrieval*.

Wang, Z.; Valdez, J.; Basu Mallick, D.; and Baraniuk, R. G. 2022. Towards Human-Like Educational Question Generation with Large Language Models. In *Proc. of Int. Conf. on Artificial Intelligence in Education*.

Welbl, J.; Liu, N. F.; and Gardner, M. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proc. of the 3rd Workshop on Noisy User-generated Text*. ACL.

Xu, H.; Van Durme, B.; and Murray, K. 2021. BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*.

Zhang, R.; Guo, J.; Chen, L.; Fan, Y.; and Cheng, X. 2021. A Review on Question Generation from Natural Language Text. *Transactions on Information Systems*, 40(1): 1–43.