CS230

Towards NLP Question-Generation that Maximizes Student Learning and Engagement

Kevin Adams
Department of Computer Science
Stanford University
kadams6@stanford.edu

Michael Herrera
Department of Computer Science
Stanford University
herrerx@stanford.edu

Abstract

Automatic question generation (AQG) for the purpose of generating assessments in educational settings is a popular subdomain within the field of Natural Language Processing [1]. However, only a handful of research projects have focused on AQG for the purpose of enhancing, rather than evaluating, students' learning [2], [3]. Research in the fields of psychology and education has shown that posing questions to students immediately after engagement with educational material improves retention [4], [5] . Research has also indicated that student engagement is critical to learning [6, pp. 40–45]. Thus, we created a novel model that, given a passage and a question, evaluates how interesting a question is (interest) and how much the question aids in comprehension of the passage (comprehension). This model could function as the discriminator component for a generative system that would generate optimally engaging and useful questions. Our model was able to achieve human-level accuracy on binary prediction of interest and comprehension.

1 Introduction

One of the main ways in which students learn new information is through reading. However, simply reading a passage once leads to poor retention of the material. Research by psychologist Henry L. Roediger has shown that quizzing oneself on information one recently learned significantly improves retention [4]. Not only does quizzing oneself lead to better retention than simply reading a passage once, it is also superior to re-reading the passage. Recall after one week is nearly two times greater when a student quizzes herself after reading rather than re-reading [5].

Existing educational material takes advantage of this phenomenon. For example, many textbooks include concept checks after each section to help students review the material they just read. Currently these concept checks are human-generated. AQG could be applied to this problem to save the time that textbook creators spend generating questions and to create concept-checks for other media (such as online articles) where it would otherwise be impractical to spend time creating questions. Making questions more interesting also improves learning outcomes. Research by the U.S. Department of Education has shown that increased student engagement leads to better learning outcomes [6, pp. 40–45].

Given the novelty of the metrics and that we had limited prior experience with NLP, we decided to focus on a classification, rather than generation, task, because it is more achievable. Thus, we created a discriminator model that, given an input of a passage and a question, produces an output binary prediction of whether the question will improve student comprehension of the passage, and whether the question will improve student interest in the topic.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

Current AQG systems are not designed to create questions that improve student comprehension of learning material. A 2019 survey of AQG systems for educational purposes identified only 7 papers focused on AQG for knowledge acquisition, and no AQG systems had evaluation metrics that considered how interesting a question is [1]. Further, an analysis of the pedagogical usefulness of question generation for reading comprehension found that there "many [evaluation] approaches focus on linguistic quality only while ignoring the pedagogic value and appropriateness of questions" [7].

Another 2019 survey paper identified that one of the key limitations to improving neural question generation was that Seq2Seq models struggle with long passages [2]. However, with newer models such as XLNet and BigBird producing strong results on passages that are thousands of tokens long, this limitation is becoming less of an issue [8], [9]. We also communicated with Yuxi Xie, who discovered that adding a relevance discriminator to question generators significantly improves human evaluation of the resulting questions [10]. Our model's architecture is inspired by her model.

3 Dataset and Features

3.1 Raw Data

Key to training our discriminator is training it on rich passages and questions. Datasets such as SQuAD would be insufficient because the passages are short and the questions are low on Bloom's taxonomy, which means that they would not be interesting to students [3]. Thus, we identified the LearningQ (LQ) dataset as the best choice because it is composed of informational passage from Khan Academy and human-generated questions [3]. LQ poses many higher-order questions that go beyond simple retrieval of information, as is the case with SQuAD [3, pp. 486–487].

The LQ dataset contains 230k questions: 7k are instructor-generated questions scraped from TED-Ed and 223k are student-generated questions scraped from Khan Academy. All questions are paired with their source documents. We selected a subset of 6k questions from the LQ dataset from Khan Academy articles in the humanities category. LQ covers various domains including mathematics, humanities, and preparation for standardized tests, but we filtered specifically for questions and articles in the humanities domain. Humanities articles tend to be more self-contained — unlike math articles, they can be understood better without the context of other articles. This aspect was critical as it allowed us to manually label the questions in our subset of LQ dataset according to our criteria of how useful and engaging a question is.

3.2 Labelling

Our model is novel in that it evaluates a question based on how engaging it is and how much it aids a student's comprehension of a preceding passage of text. Although we could get all pairs of passages and questions we needed from the LQ dataset, we still needed to label those questions based on engagement and comprehension.

We contracted Labelbox to label approximately 6246 questions, based on the ontology shown in the appendix [11]. Each crowdworker read a passage, then read the associated questions and rated them using our custom label classes shown in the appendix.

3.3 Analysis

All code for this section is located in 'analysis.ipynb'.

3.3.1 Label Distribution

88.38% of the questions were labeled understandable. The scores for comprehension were roughly bimodally distributed, with 4 being the most common label, and 2 being the second most common. Interest, on the other hand, was highly weighted 4. The label 4 constituted 65% of all entries.

Given the unequal distribution of labels, we decided to compress the labels into binary values so that there would be more entries per label. Thus, values of 1 and 2 are assigned negative binary labels and values of 4 and 5 are assigned positive binary labels. Values of 3 are discarded. We also

recognized that it would be necessary to balance the binary dataset. We explain the steps we took in the 'Processing' section below.

3.3.2 Human-level Performance

We set approximately 300 passage-question pairs to be labelled by multiple labelers. Each of these pairs was typically labelled by four people, leading to approximately 1200 total entries in our cross-validation dataset. For clarity sake, in this section, we will consider entries with the same passage-question pair to be part of the same 'group'.

We calculated human-level accuracy by performing a holdout procedure on each group in the cross-validation dataset.

1. Select one hold-out entry from the group.
2. Calculate the mean label value for the remaining entries in the group and convert it to a binary label.
3. Calculate the binary label of the hold out.
4. Compare the binary label of the hold out to the binary label calculated in step 2.
5. Repeat steps 1-5 on the remaining entries in the group.

Accuracy is calculated as the percent of times that the binary label of the hold out matched the binary label calculated in step 2. The accuracy percentage represented the accuracy with which a human labeler can correctly predict the binary label of an entry. Given this procedure, human-level prediction accuracy is 48% for comprehension, and 71% for interest. Thus, the goal of our model is to achieve these accuracy benchmarks.

3.4 Processing

Before training the model, we performed the following pre-processing steps on the dataset. The dataset used to train the interest classifier and the dataset used to train the comprehension classifier were pre-processed individually, because some of the processing steps were performed on a specific class.

1. **Average labels:** So that we could perform validation on the dataset, some passage-question pairs were labeled by multiple workers. For passage-question pairs that were labeled multiple times, we compressed them into one entry, and reculated their labels as the arithmetic mean of the group.
2. **Drop non-compressible questions:** We dropped any examples whose comprehension score was not 1.
3. **Drop entries with non-extreme labels:** Drop all entries with labels with values between 2 and 4. We dropped intermediate because we expected that only including more extreme values in the dataset would lead to more accurate binary classification.
4. **Calculate binary labels:** Values 2 or less are given a binary label of 0. Values 4 or greater are given a binary label of 1.
5. **Balance dataset:** Sample the dataset (without replacement) such that there are an equal number of positive and negative examples. Thus, the total size of the final dataset is twice the size of the binary value with the fewest examples.

4 Methods

Training entails fine-tuning a pre-trained XLNet from Huggingface [12]. We chose XLNet rather than BERT because XLNet can handle inputs greater than BERT's limit of 512 tokens, and most of our passages have between 500 and 1500 tokens [8], [13], [14].

We pre-processed our passage-question pairs using a tokenizer. The tokenizer produces three arrays: Input IDs, Attention Mask, and Token Type IDs. These tokens are fed into the model. Within the model, the token arrays are passed to the encoder. The encoder calculates embeddings for each

token and then calculates attention to model the semantic relationships within the text. The encoder produces a logit output, which is then fed through a sigmoid function to produce a logistic prediction as the model output. Figure 1 illustrates our model's architecture. Since the model output is a binary label, we chose binary cross-entropy for the loss function.

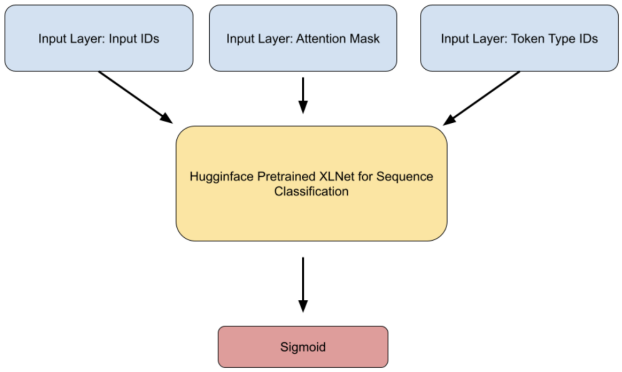

Figure 1: The model architecture

We considered creating a composite model that included this encoder and a pre-existing model (StackX) used to identify features of questions posted on StackExchange [15]. However initial analysis showed that this composite model significantly underperformed the non-composite model, and could not achieve accuracy higher than random chance. Thus, we decided to simply use the model described above. You may see references to 'stackx' in the code, but the model ignores said data.

5 Experiments/Results/Discussion

5.1 Hyperparameter Tuning

We trained the model using mini batches of size 3 because that was the maximum size that could fit in VRAM of the Tesla P100s on which we trained. Because of the small minibatch size, loss instability was a major issue. Thus, we used Adam because of its high efficacy on noisy, stochastic training problems [16]. We also searched for the optimal training rate and found that rates greater than 1e-5 lead to instability, while rates smaller than 1e-5 lead to slower training without increasing stability. Finally, we implemented an exponential decay schedule for a learning rate with a decay rate of 0.9 and decay after 100 steps. Without this decay, the loss would typically become unstable after the second epoch.

5.2 Comprehension

We were able to surpass human-level accuracy for binary classification of question comprehension. By training for 6 epochs on a set of 1,418, we achieved a training set accuracy of 68.3% and a loss of .6128. Test set accuracy was 64% with an F1 score of .64. Test set accuracy tends to level off at 65%. Training loss appears to converge after 6 epochs without improving training set performance, which indicates that our model was unable to overfit the training data (see Figure 2).

5.3 Interest

We were able to achieve human-level accuracy for binary classification of question interest. By training for 10 epochs on a set of 696 examples, we achieved training set accuracy of 92.4% and a loss of .2278. Test set accuracy was 71% with an F1 score of .71. Test set accuracy tends to level off at 70%. Training on fewer than 10 epochs leads to lower test set accuracy, but additional epochs after 10 do not provide gains in performance. Because this matches human-level accuracy, we concluded that it would be unlikely that further tuning would lead to meaningful accuracy improvements. Furthermore, loss appears to level out around epoch 9 so this further suggests that the model reaches its limit around epoch 10 (see Figure 4).

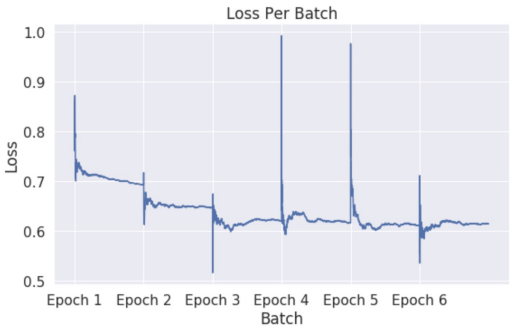

Figure 2: Comprehension training set loss.

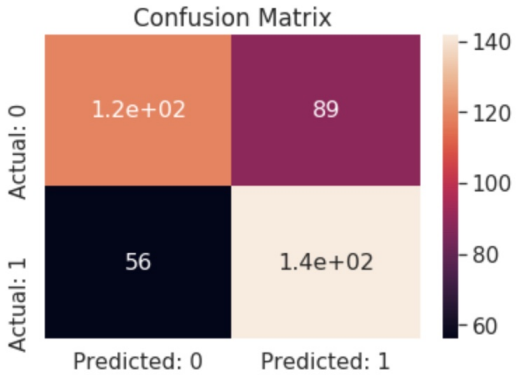

Figure 3: Comprehension test set confusion matrix

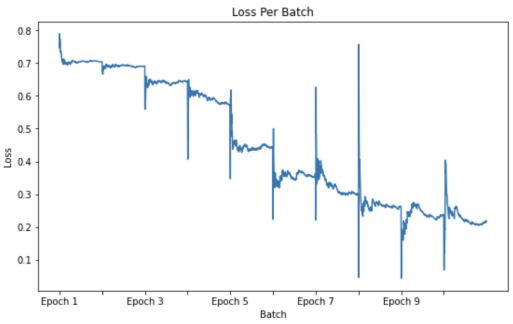

Figure 4: Training set accuracy

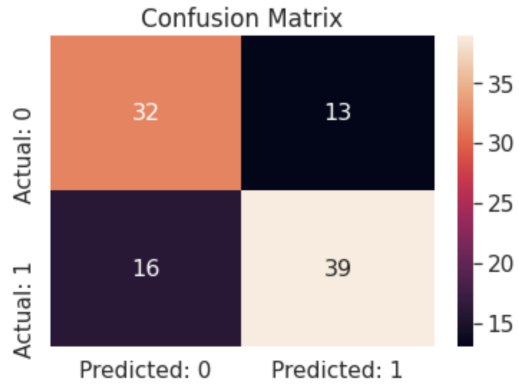

Figure 5: Training set loss

The model is highly susceptible to training set imbalance. Positive examples were over-represented in our original dataset, and caused our model to predict positive every time. However, balancing the dataset leads to balanced predictions (see Figure 3 and Figure 5).

6 Conclusion/Future Work

In this project, we fine-tuned two pre-trained XLNet models from HuggingFace: given a passage, one model classifies a question as potentially interesting to students, and the other to classifies a question as potentially helpful to students in comprehending the corresponding passage. We considered a composite approach of concatenating an XLNet model's output with the output from a pre-existing StackX model and applying a final sigmoid layer, but this approach underperformed relative to the non-composite approach.

Our final models were able to achieve at least human-level accuracy, although human-level accuracy for the comprehension task is no better than random guessing. Given that we asked labelers for subjective ratings of comprehension, the lack of consensus in comprehension labels suggests that they were more correlated to the labeler's personal preference rather than a discernible trend in question style. Our model's over 90% training accuracy in determining interest labels indicates that there is some underlying objective notion of what an interesting question is that influenced the labelers' subjective ratings. If we had more time, we would try to craft more objective comprehension labels for questions. For example, we could create comprehension-focused quizzes for each passage and observe how well each labeler performs on a quiz depending on whether they saw the corresponding questions or not. With more objective ratings for comprehension and with more computational resources to train with a larger batch size, we would hope to see less noisy results.

7 Contributions

We would like to thank Jonathan Li, our TA, in his help with this project. Thank you to Yuxi Xie for her explaining questions that we had about her AQG model. Furthermore, without Labelbox's support in labelling our dataset, this project would not have been possible.

Kevin gathered the passage-question pairs from the LearningQ dataset, worked with Labelbox to develop the ontology for human labelers, and exported the labels to generate our initial dataset. He also wrote the code to generate metrics for our model's performance.

Michael researched model architectures and wrote the python code for creating our model. He also performed data analysis on our dataset to understand the factors that affected our model's performance.

References

[1] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, no. 1, pp. 121–204, Mar. 2020, doi: 10.1007/s40593-019-00186-y.

[2] L. Pan, W. Lei, T.-S. Chua, and M.-Y. Kan, "Recent Advances in Neural Question Generation," *ArXiv190508949 Cs*, Jun. 2019, Accessed: Jan. 24, 2021. [Online]. Available: http://arxiv.org/abs/1905.08949.

[3] G. Chen, J. Yang, C. Hauff, and G.-J. Houben, "LearningQ: A Large-Scale Dataset for Educational Question Generation," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 12, no. 1, Art. no. 1, Jun. 2018, Accessed: Jan. 24, 2021. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14987.

[4] H. L. Roediger and A. C. Butler, "The critical role of retrieval practice in long-term retention," *Trends Cogn. Sci.*, vol. 15, no. 1, pp. 20–27, Jan. 2011, doi: 10.1016/j.tics.2010.09.003.

[5] H. L. Roediger and J. D. Karpicke, "Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention," *Psychol. Sci.*, vol. 17, no. 3, pp. 249–255, Mar. 2006, doi: 10.1111/j.1467-9280.2006.01693.x.

[6] F. M. Newmann, Ed., *Student engagement and achievement in American secondary schools*. New York: Teachers College Press, 1992.

[7] A. Horbach, I. Aldabe, and M. Bexte, "Linguistic Appropriateness and Pedagogic Usefulness of Reading Comprehension Questions," p. 10.

[8] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *ArXiv190608237 Cs*, Jan. 2020, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1906.08237.

[9] M. Zaheer et al., "Big Bird: Transformers for Longer Sequences," *ArXiv200714062 Cs Stat*, Jan. 2021, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/2007.14062.

[10] Y. Xie, L. Pan, D. Wang, M.-Y. Kan, and Y. Feng, "Exploring Question-Specific Rewards for Generating Deep Questions," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), 2020, pp. 2534–2546, doi: 10.18653/v1/2020.coling-main.228.

[11] "Labelbox," *Labelbox*, 2021. https://labelbox.com (accessed Feb. 26, 2021).

[12] "XLNet - transformers 4.4.1 documentation," *Hugginface*. https://huggingface.co/transformers/model_doc/xlnet.html (accessed Mar. 16, 2021).

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv181004805 Cs*, May 2019, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1810.04805.

[14] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," *ArXiv190505583 Cs*, Feb. 2020, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1905.05583.

[15] O. Yaroshevskiy, D. Danevskiy, Y. Kashnitsky, and D. Abulkhanov, *Quest QA Labeling*. 2021.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv14126980 Cs*, Jan. 2017, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1412.6980.

8 Appendix

The three custom classes of labels that human labelers assigned to each passage-question pair.

Class 1: Understandable (Boolean)

In spite of small grammatical errors, could you understand what the question was asking?

True	I could understand what this question was asking.
False	I could not understand what this question was asking.

Class 2: Comprehension (Integer, 1-5)

To what degree did the question affect your comprehension of the passage?

1	This question significantly impeded my comprehension of the passage.
2	This question moderately impeded my comprehension of the passage.
3	This question had no effect on my comprehension of the passage.
4	This question moderately improved my comprehension of the passage.
5	This question significantly improved my comprehension of the passage.

Class 3: Engagement (Integer, 1-5)

To what degree did the question affect your interest in the topic of passage?

1	This question significantly decreased my interest in the topic.
2	This question moderately decreased my interest in the topic.
3	This question had no effect on my interest in the topic.
4	This question moderately increased my interest in the topic.
5	This question significantly increased my interest in the topic.