Diverse Content Selection for Educational Question Generation

Amir Hadifar **Semere Kiros Bitew** **Johannes Deleu**
Veronique Hoste **Chris Develder** **Thomas Demesteer**
firstname.lastname@ugent.be

Abstract

Question Generation (QG) systems have shown promising results in reducing the time and effort required to create questions for students. Typically, a first step in QG is to select the content to design a question for. In an educational setting, it is crucial that the resulting questions cover the most relevant/important pieces of knowledge the student should have acquired. Yet, current QG systems either consider just a single sentence or paragraph (thus do not include a selection step), or do not consider this educational viewpoint of content selection. Aiming to fill this research gap with a solution for educational document-level QG, we thus propose to select contents for QG based on relevance and topic diversity. We demonstrate the effectiveness of our proposed content selection strategy for QG on 2 educational datasets. In our performance assessment, we also highlight limitations of existing QG evaluation metrics in light of the content selection problem.[1]

1 Introduction

Over the past few years, educational institutes have increasingly embraced digital tools and solutions to extend the purely classroom-based setting and enable wider access to high-quality education. An essential component of digital educational tools is the ability to test the learners' progress in acquiring the knowledge offered in a course. Indeed, subjecting learners to such tests triggers reflection on and consolidation of the information they have consumed (Rickards, 1979; DeAngelo et al., 2009). Yet, creating high-quality questions that comprehensively and accurately evaluate a learner's knowledge and/or skills is quite challenging due to the extensive human domain knowledge required. Current automatic Question Generation (QG) solutions have already shown significant progress in reducing the time and effort to phrase suitable questions. However, a common weakness is that they do not employ an education-oriented approach when processing full documents or book chapters (Le et al., 2014; Mostow and Chen, 2009). This implies human intervention is required to either select the input or filter suitable output questions afterwards. Note that the latter likely implies a significant computational overhead associated with the generation of questions for each and every possible paragraph/sentence.

In this research, we take one step back and focus on one of the earliest stages of developing a test called *content identification* or *content selection*, the process of reducing the amount of text in a chapter or document to its most meaningful subparts suitable for constructing questions (Kurdi et al., 2020; Davis, 2009). Content selection is a crucial and challenging step in any assessment system. It is crucial because decisions regarding which content to include or exclude can significantly influence the inferences teachers make about their students' understanding of key concepts in the considered course material. More importantly, in some settings such as self-assessment and self-learning environments, leaving the content selection to users is not feasible (Kurdi et al., 2020). It is challenging because numerous trade-offs have to be considered, such as the type of exam (e.g., low stakes vs. high stakes), the subject (e.g., mathematics vs. history), and instructor preferences, among others (Davis, 2009).

Although natural language processing has been extensively employed in educational environments (Kurdi et al., 2020; Laban et al., 2022), only a few researchers have investigated content selection for generating educational questions. Some studies (Chen et al., 2019; Rüdian et al., 2020) used summarization techniques to identify important contents. However, because these methods aim to select sentences that maximize content coverage, they may not be suitable for generating questions

[1] Code will be available at: https://github.com/hadifar/content_selection

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 123–133
May 2-4, 2023 ©2023 Association for Computational Linguistics

in the context of education, as such sentences can be incoherent and complex (Kumar et al., 2015). Steuer et al. (Steuer et al., 2021) utilized a binary classifier trained on definitions from scientific textbooks to prioritize worthy over non-worthy content. This study, however, was limited to the definition of named entities or concepts rather than general pedagogical contents. Related to our method is (Kumar et al., 2015), that ranked sentences based on topic distributions obtained from a topic model for fill-the-gaps (cloze) questions. However, unlike our proposed method, the notion of relevance of contents to teachers is ignored. Other methods (Du and Cardie, 2017; Kumar et al., 2019; Nakanishi et al., 2019; Back et al., 2021) jointly optimized content selection and QG in an end-to-end fashion. Most of the previous studies validated their methodologies by evaluating QG performance using n-gram metrics (e.g., Papineni et al. (BLEU; 2002), Banerjee and Lavie (METEOR; 2005)) instead of directly evaluating the content selection method. We will show (§4) how these metrics are inadequate to evaluate this task.

We frame *content selection* as a ranking problem that maximizes both relevance and topic diversity. The topic diversity is motivated by test development studies (Webb, 2006; Haladyna and Downing, 1989; Haladyna and Rodriguez, 2013), that suggest reliable inference regarding students' understanding of contents is tied to the number of questions that cover main topics. This hypothesis is implicitly held by teachers during question construction (Fig. 1). To this end, we propose a ranking model that assigns a score to each content (i.e., all paragraphs or sentences in a textbook), allowing us to prioritize relevant candidates. Furthermore, we introduce a re-ranker that encourages topic diversity. Our empirical results (§4) not only show that our model leads to an improved content ranking compared to existing methods on two recently released educational datasets but also reveal difficulties in measuring ranking quality through evaluation of the question generation end task.

2 Methodology

This section describes our strategy for obtaining a suitable ranking of a document's sentences or paragraphs. Since optimizing relevance and diversity at the same time is an NP-hard problem (Agrawal et al., 2009), heuristic approaches are typically used (Carbonell and Goldstein, 1998; Santos et al., 2010), which we propose for our setting as well. After an initial ranking purely based on estimated relevance, we apply an iterative diversity-aware reranking.

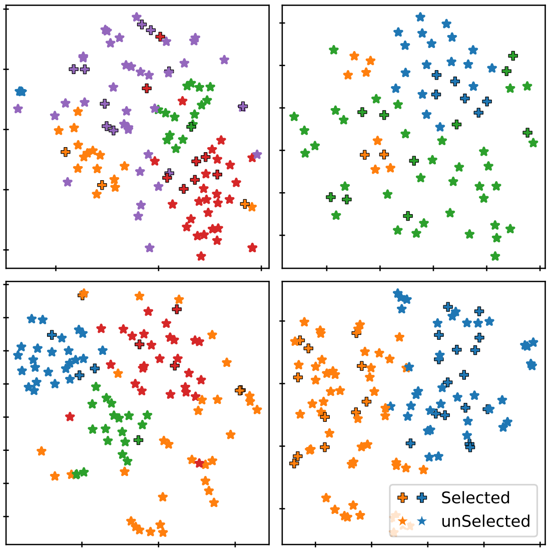

Figure 1: 2D visualization of the pooled hidden-state representations of the RoBERTa (Liu et al., 2019) on all paragraphs for four randomly sampled chapters of EduQG (Hadifar et al., 2023) using t-SNE (van der Maaten and Hinton, 2008). Each color stands for a different topic assigned by the Gaussian clustering model, while the marker types represent the selected vs. unselected paragraphs by teachers.

Relevance-based ranking: We assume that the considered educational content involves textual documents D (e.g., course book chapters), each represented as a set of N content elements $D = \{s_0, s_1, s_2, ..., s_{N-1}\}$. These elements s_i can be, for example, the document's sentences or paragraphs (and will be referred to as the sentences). In our datasets, an associated question is available for a selection of the sentences, created by a teacher. Although different teachers would likely not agree on which sentences are relevant (i.e., to create questions from) given the same document, we consider the sentences associated with the available questions as a good proxy. As such, we train a classifier to predict for each sentence whether it is relevant or not. In our experiments, this is achieved by training a logistic regression classifier on top of a pre-trained language model's representation (Liu et al., 2019), similar to the approach in Nogueira and Cho (2019). At inference, after ranking all sentences according to decreasing relevance scores as predicted by that model, each sentence s_i is represented by the *relevance score* $R_i = 1 - \rho_i/N$, with ρ_i its rank (i.e., the highest ranked sentence is scored 1, and the lowest one $1/N$).

Diversity-aware reranking: Next, all sentences s_i are iteratively reranked by combining the relevance-based score R_i with a score that promotes *topic-related diversity*, in line with educational insights (Davis, 2009). The following paragraphs outline this procedure.

First, topics are identified through a Gaussian clustering model, fit to all (language model-based) sentence representations $x_i = RoBERTa(s_i)$. The likelihood $p(s_i)$ of a sentence s_i over all topics is written as $p(s_i) = \sum_z p(z)\, p(s_i|z)$, with the topic probability $p(z)$ and the gaussian mixture component $p(s_i|z) = \mathcal{N}(\mathbf{x}_i; \mu_z, \Sigma_z)$ (with mean vector μ_z and covariance matrix Σ_z).After obtaining a fit for the topic probabilities and gaussian component parameters, the topic distribution for each sentence can be readily obtained as $p(z|s_i) = p(s_i, z)/\sum_{z'} p(s_i, z')$. Next, we initialize the final sentence ranking S with the sentence that received the highest relevance-oriented rank R_i. We then iteratively add sentence by sentence, by combining their relevance score and a diversity score that measures topic diversity with respect to the sentences already present in S. During every considered iteration, the diversity scores $D(s_i|S)$ are calculated as follows, for any sentence $s_i \notin S$:

$$D(s_i|S) = \sum_z p(z)\Big(p(s_i, z) \prod_{s_j \in S} (1 - p(s_j, z))\Big)$$

which can be interpreted as the expectation over all topics, that a given topic would occur with s_i *and* with none of the sentences already ranked in S (if the considered sentences were independent, strictly speaking). The quantities $p(s_i, z)$ are approximated as $p(s_i, z) \approx R_i\, p(z|s_i)$, substituting the prior probability of sentence s_i by the relevance score R_i. For each sentence s_i not yet ranked in S, the relevance and diversity scores are combined into the score $\lambda R_i + (1-\lambda) D(s_i|S)$, with a weighting coefficient λ. The highest scoring sentence is then added to S, and the next iteration starts.

The method described above is inspired by a well-known technique for query reformulation for web search results diversification (Santos et al., 2010). It would likely work with alternative topic models as well. Note that in our experiments, we do not predefine the number of topics, which is estimated through the bayesian information criterion (Schwarz, 1978). However, a teacher could alter the number of topics manually, depending on the desired level of granularity in the topics.

3 Experimental Setup

Datasets. Most existing QG datasets are neither educational (e.g., SQuAD (Rajpurkar et al., 2016)) or do not provide an explicit link between questions and course content (e.g., LearningQ (Chen et al., 2018)) making it impossible to evaluate content selection methods directly. To the best of our knowledge, the only educational datasets that allow for such evaluation are EduQG (Hadifar et al., 2023) and TQA-A (Steuer et al., 2022). EduQG contains questions (i.e., phrased in cloze or close-ended form) and correct answers that are sentence-level aligned to source documents. TQA-A contains question-answer pairs where answers are annotated at the span level. We evaluate our content selection methods at the paragraph and sentence levels, respectively, for EduQG and TQA-A (an example entry of each dataset is presented in Appendix §A).

Baselines. We compare different content selection baselines including: ORACLE (a perfect retrieval system), LEXRANK (Erkan and Radev, 2004), SVM (Cortes and Vapnik, 1995), *Overgeneration and Rank* (OVR; Heilman and Smith, 2009). The baselines are compared against OUR methodology (§2) for some selected λ values. The T5 model (Raffel et al., 2020) has been fine-tuned to function as the QG model for the baselines. We devised a fixed QG model for all content selection strategies in each dataset, in order to obtain a fair evaluation regarding diversity and generation quality (see Appendix §B for further details).

Evaluation. To measure the content ranking performance, we report Recall (R) and Mean Average Precision (MAP) for the top 10 candidates. Diversity, for the selected sentences/paragraphs as well as among the generated questions, is measured by Average cosine Distance between Candidates (ADC; Belém et al., 2013), Self-BLEU (SBLEU; Zhu et al., 2018), and distinct-unigram (DIST1; Li et al., 2015). We also report BLEU and METEOR to measure the quality of the generated questions (more details in Appendix §C).

Table 1: Summary of our results on EduQG and TQA-A datasets.

		Retrieval		Content-diversity			Question-diversity			Generation	
	Method	**R**	**MAP**	**ADC**	**SBLEU↓$^{(*)}$**	**DIST1**	**ADC**	**SBLEU↓**	**DIST1**	**BLEU**	**METEOR**
EduQG	Oracle	100	100	67.5	7.8	37.1	65.8	49.0	36.3	35.6	46.7
	LexRank	20.0	37.2	50.7	13.9	34.4	54.6	54.4	31.8	33.3	45.0
	SVM	28.5	46.3	63.9	9.2	32.8	66.4	50.8	35.7	33.3	45.4
	OvR	-	-	-	-	-	67.1	**48.0**	36.9	31.0	39.8
	Our (λ=1.0)	**32.7**	**51.6**	64.5	8.7	33.8	66.5	49.2	37.0	**33.7**	**45.8**
	Our (λ=0.01)	23.6	40.3	75.2	7.1	38.5	**73.6**	48.6	**37.8**	31.8	43.3
	Our (λ=0.0)	12.2	43.2	**76.1**	**6.0**	**44.1**	70.6	**46.7**	37.3	31.8	42.5
TQA-A	Oracle	100	100	57.6	66.1	40.3	48.2	65.4	49.1	7.4	19.4
	LexRank	20.7	31.7	51.7	74.1	35.4	45.5	69.1	46.5	7.1	**19.5**
	SVM	20.2	32.6	56.7	72.2	38.5	48.7	67.8	47.3	7.8	18.6
	OvR	-	-	-	-	-	**70.2**	**26.5**	**79.9**	6.0	17.3
	Our (λ=1.0)	**26.2**	**39.7**	57.3	63.7	41.0	51.3	57.5	53.2	**8.2**	19.1
	Our (λ=0.01)	22.9	38.0	**61.7**	49.2	45.3	64.3	34.1	60.6	7.6	19.2
	Our (λ=0.0)	19.6	32.9	60.8	**42.9**	**48.5**	61.2	37.2	59.2	7.1	18.6

($*$) Lower is better as it indicates higher diversity.

4 Results and Discussion

Table 1 presents a summary of our results on both datasets. The transformer-based relevancy estimator, Our ($\lambda = 1.0$), obtains the highest retrieval scores. However, this high recall or MAP does not neccessary translate into better BLEU or METEOR score (see column 'Generation'). For example, LexRank leads to almost the same overall QG quality. This phenomenon was already present in previous studies (Chen et al., 2019; Mahdavi et al., 2020), although not explicitly addressed. In addition, not even the generation scores based on perfect rankings (Oracle) are consistently better than those from predicted rankings (See Appendix E for some generated examples). This implies that the current QG evaluation metrics are incapable of evaluating the content selection step, given the QG quality of present-day competitive models like our tuned T5, and these two tasks must be evaluated separately. Alternatively, asking experts to review the quality of generated questions or content selection as done in previous studies (Steuer et al., 2022; Huang and He, 2016) is not reproducible.

We can see a clear correlation between Content-diversity and Question-diversity columns. For instance, Our ($\lambda = 0.0$) selection leads to the highest content diversity and, consequently, the highest question diversity. A higher degree of diversity can be obtained by decreasing λ, at the expense of retrieval effectiveness. The correlation between content diversity and question diversity further supports our suggestion to split up the evaluation of content selection and question generation.

The OvR strategy gets better diversity scores on TQA-A. We observed the ranker prioritize cloze questions over "wh" questions. Our hypothesis is that the lower word overlap on "wh" words in the highest ranked questions leads to the observed higher diversity. In EduQG, we do not observe this behavior, since all questions have the same format. In fact, We believe that selecting content first and then generating questions is a better strategy compared to OvR. This aligns more closely with how teachers typically create questions and is also more computationally efficient.[2], and in terms of quality of the resulting question set (cf. higher generation scores for Our models than OvR).

Summarized, our experiments lead to the following insights: (i) Our proposed model is able to outperform all baselines in terms of retrieval metrics. (ii) It allows control over the trade-off between retrieval quality and diversity (through the parameter λ). (iii) Content-diversity and question-diversity do correlate (which is less than surprising), but neither retrieval nor diversity seems to correlate well with established metrics to evaluate generated questions.

5 Conclusion

This paper describes an educational-oriented strategy for content selection from educational documents to support question generation, using a ranker and a topic-wise diversifier. Our empirical

[2] Note that when using content selection, we only generate a limited set of questions, rather than all possible ones for a chapter, as we do with the OvR strategy. Based on our analysis, the inference time for question generation using T5-base is an order of magnitude higher than ranking with RoBERTa-base (288.69 vs. 14.29 milliseconds for a single inference pass).

evaluations of two educational QG datasets demonstrate the effectiveness of the proposed model. However, we found that current ngram-based evaluation metrics of the generated questions, given the current level of generation quality, do not carry sufficient signal to evaluate the content selection problem.

Limitations

We believe the current study can improve in at least two ways: (i) The limitations of existing QG evaluation metrics in light of the content selection problem are highlighted in this study, however, a promising next step is to annotate topics/subtopics and evaluate the diversity of generated contents and questions by more sophisticated metrics such as α-NDCG (Clarke et al., 2008) or ERR-AI (Chapelle et al., 2009). (ii) A human study on content selection and question generation will be insightful. For example, the analysis of question diversity and its impact on students' learning allows us to understand the necessity of diversification better.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of WSDM*.

Seohyun Back, Akhil Kedia, Sai Chetan Chinthakindi, Haejun Lee, and Jaegul Choo. 2021. Learning to generate questions by learning to recover answer-containing sentences. In *Proceedings of ACL-IJCNLP*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of WS*.

Fabiano Belém, Eder Martins, Jussara Almeida, and Marcos Gonçalves. 2013. Exploiting novelty and diversity in tag recommendation. In *Proceedings of ECIR*, pages 380–391.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the SIGIR*.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of CIKM*.

Guanliang Chen, Jie Yang, and Dragan Gasevic. 2019. A comparative study on question-worthy sentence selection strategies for educational question generation. In *Proceedings of AIED*.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. LearningQ: a large-scale dataset for educational question generation. In *Proceedings of AAAI*.

Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3).

Barbara Gross Davis. 2009. *Tools for teaching*. John Wiley & Sons.

Linda DeAngelo, Sylvia Hurtado, John H Pryor, Kimberly R Kelly, Jose Luis Santos, and William S Korn. 2009. The american college teacher: National norms for the 2007-2008 HERI faculty survey. *Los Angeles: Higher Education Research Institute, UCLA*.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of EMNLP*.

Günes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. EduQG: A multi-format multiple choice dataset for the educational domain. *IEEE Access*.

Thomas M Haladyna and Steven M Downing. 1989. A taxonomy of multiple-choice item-writing rules. *Applied measurement in education*, 2(1).

Thomas M Haladyna and Michael C Rodriguez. 2013. *Developing and validating test items*. Routledge.

Michael Heilman and Noah A Smith. 2009. Question generation via overgenerating transformations and ranking. Technical report, Carnegie-Mellon University Pittsburgh - language technologies institute.

Yan Huang and Lianzhen He. 2016. Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, 22(3):457–489.

Girish Kumar, Rafael E Banchs, and Luis Fernando D'Haro. 2015. Revup: Automatic gap-fill question generation from educational texts. In *Proceedings of SIGEDU*.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2019. Putting the horse before the cart: A generator-evaluator framework for question generation from text. In *Proceedings of CoNLL*.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1).

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz design task: Helping teachers create quizzes with automated question generation. In *Proceedings of NAACL*.

Nguyen-Thinh Le, Tomoko Kojiri, and Niels Pinkwart. 2014. Automatic question generation for educational applications–the state of art. In *Proceedings of ICCSAMA*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sedigheh Mahdavi, Aijun An, Heidar Davoudi, Marjan Delpisheh, and Emad Gohari. 2020. Question-worthy sentence selection for question generation. In *Proceedings of CANAI*.

Jack Mostow and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of AIED*.

Mao Nakanishi, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2019. Towards answer-unaware conversational question generation. In *Proceedings of MRQA*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.

John P Rickards. 1979. Adjunct postquestions in text: A critical review of methods and processes. *Review of Educational Research*, 49(2).

Sylvio Rüdian, Alexander Heuts, and Niels Pinkwart. 2020. Educational text summarizer: Which sentences are worth asking for? *Fachtagung Bildungstechnologien der Gesellschaft für Informatik*.

Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW*.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics*.

Tim Steuer, Anna Filighera, Tobias Meuser, and Christoph Rensing. 2021. I do not understand what I cannot define: Automatic question generation with pedagogically-driven content selection. *arXiv preprint arXiv:2110.04123*.

Tim Steuer, Anna Filighera, and Thomas Tregel. 2022. Investigating educational and noneducational answer selection for educational question generation. *IEEE Access*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9.

Norman L Webb. 2006. Identifying content for student achievement tests. *Handbook of test development*.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems*, 40(1).

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. TexyGen: A benchmarking platform for text generation models. In *Proceedings of SIGIR*.

Appendices

A Datasets

An example of a randomly selected chapter and corresponding question(s) for EduQG and TQA-A is presented in Table 3 and Table 4 respectively. As can be seen in the tables, the length of the chapter in TQA-A is significantly shorter than in EduQG.

B Baselines

In this section, we provide more details about our baselines presented in §3. As mentioned, we compare different baselines including: (i) ORACLE: a perfect retrieval system that simulates teachers' behaviors for selecting suitable sentences or paragraphs. We operated under the assumption that the ORACLE is flawless and exactly retrieves the same number of content as teachers. (ii) LEXRANK (Erkan and Radev, 2004): the graph-based automatic text summarizer from (Chen et al., 2019),[3] (iii) SVM (Cortes and Vapnik, 1995): a pointwise ranking that is comparable to the feature-based strategy from (Mahdavi et al., 2020).[4] We fine-tuned the T5-base (Raffel et al., 2020) (see Appendix §D for the configurations) in answer-agnostic mode (Zhang et al., 2021) on our datasets and used it as the QG model for all of our content selection baselines. The greedy decoding is employed because other strategies have minimal impact on final results.

As a final baseline, we apply the fundamentally different strategy called overgeneration, a highly popular technique in the educational QG literature (Kurdi et al., 2020). First, all possible questions are generated ('overgeneration') and ranked, after which the exam designers decide which questions to select. Similar to *Overgeneration and Rank* (OvR) (Heilman and Smith, 2009), we generated all possible questions in a chapter and trained a ranker to select the most important questions. However, we replaced their rule-based QG and linear regression ranker with our T5-based QG model and new RoBERTa-base question ranker for a fair comparison with the other baselines. We utilized a similar setup for sorting questions (pointwise-ranker with cross-entropy loss). The above baselines are compared against Our methodology (§2) for various values of λ.

C Evaluation metrics

To evaluate different retrieval strategies, we feed the selected contents (sentences or paragraphs) to the finetuned T5 and evaluate the effectiveness of strategies with automatic diversity and quality metrics. The diversity of the selected sentences/paragraphs, as well as among the generated questions is measured[5] by Average cosine Distance between Candidates (ADC) (Belém et al., 2013), Self-BLEU (SBLEU) (Zhu et al., 2018), and distinct-ngram (DIST1, $n = 1$ in our case) (Li et al., 2015). ADC calculates the average cosine distance dissimilarity between the representations of all pairs. Similar to ADC, SBLEU, computes the average BLEU score (Papineni et al., 2002) between one instance and others by considering the instance as a hypothesis and the other as references. The lower SBLEU score indicates the higher diversity. Distinct ngram computes the proportion of unique n-grams out of the total number of n-grams in a set of generated questions.

The generation quality assessed by BLEU and METEOR scores.[6] BLEU relies on the maximum n-grams for counting the co-occurrences between the generated question by the generative model (i.e., T5), and a set of ground truth reference questions constructed by a teacher. The final score is derived from the average of BLEU scores through all examples. METEOR is calculated similarly by considering stemming and synonymy into account.

D Hyperparameters

Both datasets comes with a predefined train-test split. For all tasks (i.e., ranking and generation), we hold out 10% of the data for validation, while the remaining part is used for training. T5 was fine-tuned separately for both datasets from pretrained 'base' version[7] with the following hyperparameter settings in an answer agnostic[8] way:

```
batch_size=8
total_epochs=10
max_source_length=512
max_target_length=64
optimizer=AdamW
weight_decay=0.1
adam_epsilon=1e-08
max_grad_norm=1.0
lr_scheduler=linear
learning_rate=5e-05
warmup_steps=500
gradient_accumulation_steps=4
```

The presented λ for Our methodology in Table 1 selected carefully to illustrate its effect across the datasets. Figure 2 shows different values of lambda and corresponding ADC metrics in EduQG and TQA-A.

E Example Generations

In this section, we provide some examples to illustrate the limitation of the existing metrics to evaluate content selection methods. Table 2 provides a set of generated questions based on different retrieval strategies presented in §4 for a chapter in

[3] https://github.com/crabcamp/lexrank

[4] https://scikit-learn.org

[5] α-NDCG (Clarke et al., 2008) or other popular metrics for retrieval diversity have not been reported, by lack of ground-truth topic annotations.

[6] We used an existing implementation from https://www.nltk.org/

[7] https://huggingface.co/docs/transformers

[8] Note that assuming access to the answer during inference time is not always a valid option in real-world applications.

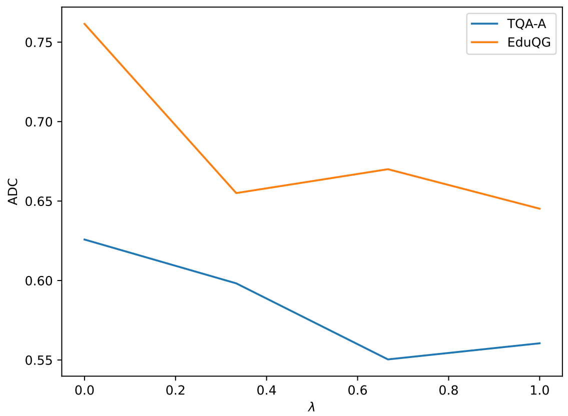

Figure 2: The effect of λ values on diversity of contents (in terms of the Average Distance between Candidate (ADC) metric).

the EduQG dataset. In addition to generated questions based on different retrieval strategies, we also show questions that were originally constructed by a teacher for the chapter (denoted by REFERENCE). It also should be noted that the difference between REFERENCE and ORACLE arises from the fact that generated questions from ORACLE were produced by the T5 model rather than a teacher. We also report BLEU and METEOR scores for these strategies.

As can be seen in the table, the top two performing retrieval systems, ORACLE and OUR $(\lambda = 1.0)$, for the EduQG dataset in our experiments (§4) lead to the lower BLEU scores compared to the weaker baselines (in terms of MAP score) such as SVM or OUR $(\lambda = 0.0)$. This issue arises from the fact that the BLEU score heavily penalizes examples that have no tri-gram or 4-gram overlap. As an example, the last two questions in ORACLE blocks (**Q5** and **Q6**) obtained BLEU scores of 0.76 and 0.63, respectively (almost zero). Although these questions seem reasonable, the last two questions in SVM block received much better scores, 25.96 and 30.21, due to the 4-gram overlap with REFERENCE set. Therefore, ORACLE retrieval receives a lower score compared to SVM despite its perfect recall and MAP.

Method	Questions	BLEU	METEOR
REFERENCE	**Q1** The study of nutrient cycling through the environment is an example of which of the following? **Q2** Understory plants in a temperate forest have adaptations to capture limited which of the following? **Q3** Which of the following biomes is characterized by abundant water resources? **Q4** Which of the following biomes is characterized by short growing seasons? **Q5** What is a key feature of estuaries? **Q6** Which of the following natural forces is responsible for the release of carbon dioxide and other atmospheric gases?	100.0	100.0
ORACLE	**Q1** Which of the following is correct about the Arctic tundra? **Q2** Which of the following is a natural driver of climate change? **Q3** What are plants that grow on other plants that are not harmed? **Q4** Which of the following is an example of a barren habitat? **Q5** What is the ecosystem composed of? **Q6** What is the seasonality of tropical wet forests?	18.92	29.23
LEXRANK	**Q1** Which of the following is not a characteristic of freshwater biomes? **Q2** What are freshwater biomes? **Q3** Water is a source of drinking water for the city. **Q4** Which of the following is correct about the spring and fall turnover? **Q5** What happens to the lake's surface water when it cools to 4 degrees C? **Q6** Climate change is a term used to describe changes in weather patterns that have become increasingly evident over	10.52	27.74
SVM	**Q1** Which of the following is not found in the neritic zone? **Q2** Which of the following is correct about the spring and fall turnover? **Q3** Which of the following is correct about the intertidal zone? **Q4** Which of the following is correct about the Little Ice Age? **Q5** Which of the following is not a characteristic of the chaparral? **Q6** Which of the following is correct about temperate grasslands?	25.56	37.94
OUR ($\lambda = 1.0$)	**Q1** Which of the following is correct about the Challenger Deep? **Q2** Which of the following is correct about the intertidal zone? **Q2** What are environments in which the soil is permanently or periodically saturated with water? **Q4** What is the amount of organic matter available as food called? **Q5** Which of the following is correct about the deepest part of the ocean? **Q6** Which of the following is not a characteristic of the deepwater region of the ocean?	16.21	32.23
OUR ($\lambda = 0.01$)	**Q1** Which of the following is correct about the Challenger Deep? **Q2** In which of the following regions would you expect to find photosynthetic organisms? **Q3** Which of the following is correct about the Milankovitch cycles? **Q4** Which of the following is correct about subtropical deserts? **Q5** What are lakes and ponds? **Q6** Which of the following is an endemic species?	23.25	33.51
OUR ($\lambda = 0.0$)	**Q1** Which of the following is correct about the Challenger Deep? **Q2** Which of the following is correct about abiotic forces? **Q3** Which of the following is not a marine biome? **Q4** Which of the following is correct about the environment? **Q5** What is the net primary productivity of boreal forests? **Q6** Which of the following is correct about coral reefs?	25.55	39.75

Table 2: A set of generated questions based on different retrieval strategies for a cherry-picked chapter in the EduQG dataset.

[Begining of the chapter is truncated due to the length limit]

Socialism is an alternative economic system. In socialist societies, the means of generating wealth, such as factories, large farms, and banks, are owned by the government and not by private individuals. The government accumulates wealth and then redistributes it to citizens, primarily in the form of social programs that provide such things as free or inexpensive health care, education, and childcare. In socialist countries, the government also usually owns and controls utilities such as electricity, transportation systems like airlines and railroads, and telecommunications systems. In many socialist countries the government is an oligarchy : only members of a certain political party or ruling elite can participate in government. For example, in China, the government is run by members of the Chinese Communist Party. However, socialist countries can have democratic forms of government as well, such as Sweden. Although many Americans associate socialism with tyranny and a loss of individual liberties, this does not have to be the case, as we see in Sweden.

In the United States, the democratic government works closely together with its capitalist economic system. The interconnectedness of the two affects the way in which goods and services are distributed. The market provides many goods and services needed by Americans. For example, food, clothing, and housing are provided in ample supply by private businesses that earn a profit in return. These goods and services are known as private goods . 1 People can purchase what they need in the quantity in which they need it. This, of course, is the ideal. In reality, those who live in poverty cannot always afford to buy ample food and clothing to meet their needs, or the food and clothing that they can afford to buy in abundance is of inferior quality. Also, it is often difficult to find adequate housing; housing in the most desirable neighborhoods—those that have low crime rates and good schools—is often too expensive for poor or working-class (and sometimes middle-class) people to buy or rent.

Thus, the market cannot provide everything (in enough quantity or at low enough costs) in order to meet everyone's needs. Therefore, some goods are provided by the government. Such goods or services that are available to all without charge are called public goods. Two such public goods are national security and education. It is difficult to see how a private business could protect the United States from attack. How could it build its own armies and create plans for defense and attack? Who would pay the men and women who served? Where would the intelligence come from? Due to its ability to tax, draw upon the resources of an entire nation, and compel citizen compliance, only government is capable of protecting the nation.

Similarly, public schools provide education for all children in the United States. Children of all religions, races and ethnicities, socioeconomic classes, and levels of academic ability can attend public schools free of charge from kindergarten through the twelfth grade. It would be impossible for private schools to provide an education for all of the nation's children. Private schools do provide some education in the United States; however, they charge tuition, and only those parents who can afford to pay their fees (or whose children gain a scholarship) can attend these institutions. Some schools charge very high tuition, the equivalent to the tuition at a private college. If private schools were the only educational institutions, most poor and working-class children and many middle-class children would be uneducated. Private schooling is a type of good called a toll good . Toll goods are available to many people, and many people can make use of them, but only if they can pay the price. They occupy a middle ground between public and private goods. All parents may send their children to public schools in the United States. They can choose to send their children to a private school, but the private school will charge them. On the other hand, public schools, which are operated by the government, provide free education so all children can attend school. Therefore, everyone in the nation benefits from the educated voters and workers produced by the public school system. Another distinction between public and private goods is that public goods are available to all, typically without additional charge.

[Rest of the chapter is truncated due to the length limit]

Reference Question:
Q1 What goods are available to all without direct payment?
a) private goods b) public goods c) common goods d) toll goods

[Rest of the questions are truncated due to the length limit]

Table 3: An example of a randomly selected chapter in the EduQG dataset with a reference question. The highlighted paragraph indicates the selected content or the grounding answer.

Dust storms like the one in Figure 10.20 are more common in dry climates. The soil is dried out and dusty. Plants may be few and far between. Dry, bare soil is more easily blown away by the wind than wetter soil or soil held in place by plant roots.
Like flowing water, wind picks up and transports particles. Wind carries particles of different sizes in the same ways that water carries them. You can see this in Figure 10.21. Tiny particles, such as clay and silt, move by suspension. They hang in the air, sometimes for days. They may be carried great distances and rise high above the ground. Larger particles, such as sand, move by saltation. The wind blows them in short hops. They stay close to the ground. Particles larger than sand move by traction. The wind rolls or pushes them over the surface. They stay on the ground. Did you ever see workers sandblasting a building to clean it? Sand is blown onto the surface to scour away dirt and debris. Wind-blown sand has the same effect. It scours and polishes rocks and other surfaces. Wind-blown sand may carve rocks into interesting shapes. You can see an example in Figure 10.22. This form of erosion is called abrasion. It occurs any time rough sediments are blown or dragged over surfaces. Can you think of other ways abrasion might occur? Like water, when wind slows down it drops the sediment its carrying. This often happens when the wind has to move over or around an obstacle. A rock or tree may cause wind to slow down. As the wind slows, it deposits the largest particles first. Different types of deposits form depending on the size of the particles deposited. When the wind deposits sand, it forms small hills of sand. These hills are called sand dunes. For sand dunes to form, there must be plenty of sand and wind. Sand dunes are found mainly in deserts and on beaches. You can see examples of sand dunes in Figure 10.23. What causes a sand dune to form? It starts with an obstacle, such as a rock. The obstacle causes the wind to slow down. The wind then drops some of its sand. As more sand is deposited, the dune gets bigger. The dune becomes the obstacle that slows the wind and causes it to drop its sand. The hill takes on the typical shape of a sand dune, shown in Figure 10.24. Once a sand dune forms, it may slowly migrate over the land. The wind moves grains of sand up the gently sloping side of the dune. This is done by saltation. When the sand grains reach the top of the dune, they slip down the steeper side. The grains are pulled by gravity. The constant movement of sand up and over the dune causes the dune to move along the ground. It always moves in the same direction that the wind usually blows. Can you explain why? When the wind drops fine particles of silt and clay, it forms deposits called loess. Loess deposits form vertical cliffs. Loess can become a thick, rich soil. Thats why loess deposits are used for farming in many parts of the world. You can see an example of loess in Figure 10.25. Its very important to control wind erosion of soil. Good soil is a precious resource that takes a long time to form. Covering soil with plants is one way to reduce wind erosion. Plants and their roots help hold the soil in place. They also help the soil retain water so it is less likely to blow away. Planting rows of trees around fields is another way to reduce wind erosion. The trees slow down the wind, so it doesnt cause as much erosion. Fences like the one in Figure 10.26 serve the same purpose. The fence in the figure is preventing erosion and migration of sand dunes on a beach.

Reference Questions:
Q1 Wind drops the sediment it is carrying when it ...
a) slows down. b) is very moist. c) arrives at a beach. d) reaches a certain altitude.

Q2 A sand dune migrates because wind keeps
a) reversing its direction. b) blowing sand up and over the dune. c) causing longshore drift. d) none of the above

Q3 Deposits called loess
a) form vertical cliffs. b) have thick rich soil. c) are deposited by wind. d) all of the above

Q4 Loess deposits consist of
a) sand and silt. b) silt and clay. c) clay and gravel. d) gravel and sand.

Table 4: An example of a randomly selected chapter in the TQA-A dataset with reference questions. The highlighted sentences indicate the selected contents or the grounding answers.