ConQuest: Contextual Question Paraphrasing through Answer-Aware Synthetic Question Generation

Mostafa Mirshekari[1] **Jing Gu**[1, 2] **Aaron Sisto**[1]
[1] Searchable.ai, USA
[2] University of California, Santa Cruz, USA
{mostafa,aaron}@searchable.ai
jgu110@ucsc.edu

Abstract

Despite excellent performance on tasks such as question answering, Transformer-based architectures remain sensitive to syntactic and contextual ambiguities. Question Paraphrasing (QP) offers a promising solution as a means to augment existing datasets. The main challenges of current QP models include lack of training data and difficulty in generating diverse and natural questions. In this paper, we present *ConQuest*, a framework for generating synthetic datasets for contextual question paraphrasing. To this end, *ConQuest* first employs an answer-aware question generation (QG) model to create a question-pair dataset and then uses this data to train a contextualized question paraphrasing model. We extensively evaluate *ConQuest* and show its ability to produce more diverse and fluent question pairs than existing approaches. Our contextual paraphrase model also establishes a strong baseline for end-to-end contextual paraphrasing. Further, We find that context can improve BLEU-1 score on contextual compression and expansion by 4.3 and 11.2 respectively, compared to a non-contextual model.

1 Introduction

In recent years, Transformer-based architectures have made enormous progress in neural question answering (QA) (Karpukhin et al., 2020) (Izacard and Grave, 2020). However, these models are still sensitive to query syntax and ambiguity (Buck et al., 2017) (Moon and Fan, 2020). While data augmentation and lexical normalization are popular approaches to account for syntactic variations of input texts, progress on generating diverse and meaningful question variants (i.e., question paraphrasing (QP)) has been limited.

The main challenge in QP remains the lack of large-scale question paraphrase datasets. One potential solution is to use a QG model to generate multiple questions based on a given context and use these as paraphrased versions of each other. The challenge with this approach is the difficulty in generating questions that are diverse, fluent, and consistent. Further, these questions should be of different lengths to make sure that the QA models learn which information in the question to consider or neglect. This can be achieved by considering question expansion/compression tasks. However, in order to generate such questions, the context surrounding the source question and answer becomes important. The expansion task requires associating sparse information in a short-form question with details from the context to reformulate a richer question. Conversely, compression requires synthesizing extraneous details in a long-form question into a shorter form that is still consistent with the given passage and answer.

Here, we introduce *ConQuest*, a framework for generating synthetic contextual question paraphrase datasets for tasks such as question compression and expansion. Specifically, given a passage as context, we first employ an answer-aware Sequence-to-Sequence (Seq2Seq) model to generate a diverse set of question variants, each consistent with a common answer span. Then, we pair the shortest and longest variants to form a novel contextual paraphrasing dataset suitable for question expansion and compression tasks. Being answer-aware, this QG model accounts for contextual information when linking question pairs (unlike existing paraphrase methods using back-translation (Xie et al., 2019) or term replacement (Mrksic et al., 2016)).

Using this synthetic dataset, the final module of our framework is a Seq2Seq model for contextual question expansion and compression. By considering the context, this module addresses the unexplored challenge of contextual question reformulation. Further, we describe a multi-tag encoding scheme for compression and expansion to improve the quality of generated questions of dif-

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 222–229
November 11, 2021.

ferent length by providing an effective method to disentangle their representations while leveraging their shared properties. To evaluate our framework, we use questions from SQuAD v1.1 (Rajpurkar et al., 2016), a common question answering dataset. We measure performance of each component in our framework across a number of automatic and human-based metrics. We will release our question paraphrasing dataset and code upon acceptance.

2 Related Work

In this section, we review previous results and challenges in question generation and reformulation, and discuss novel aspects of our contextual question paraphrasing dataset.

2.1 Answer-Aware Question Generation

Answer-aware question generation (QG) involves encoding a source passage alongside the target answer to generate a consistent, fluent question. Seq2seq models in particular have been explored for this task (Liu et al., 2019; Kim et al., 2019; Ma et al., 2020; Varanasi et al., 2020; Rajpurkar et al., 2016; Gu et al., 2021; Majumder et al., 2021). Although QG performance has increased on the inverted SQuAD task, the diversity and fluency of generated questions for paraphrasing has not been studied quantitatively.

2.2 Comparison to Existing Datasets

Open source question paraphrase datasets are uncommon due to the private nature of user queries on most search engines. One relevant dataset is CANARD (Elgohary et al., 2019), which is a conversational QP dataset based on CoQA (Reddy et al., 2018). However, CANARD is not suitable for tasks like contextual compression and expansion as it focuses primarily on contextual coreference alignment and syntactic variations. Other datasets containing question paraphrases, like PAWS-X (Yang et al., 2019) and ComQA (Abujabal et al., 2019), are more appropriate for paraphrase identification and non-contextual paraphrase generation tasks. Our synthetic dataset expands on existing work by providing both context and question pairs to enable contextual paraphrasing and generating question pairs with diverse lengths to enable expansion and compression.

3 Method

In this section, we describe *ConQuest*, our contextual question paraphrasing framework. *ConQuest* has two main modules: 1) Paraphrase Question Generation where answer-aware question generation is used to create a question-pair dataset and 2) Contextual Paraphrasing to train a generative model to compress and expand these questions. These two modules and their components are shown in Figure 1.

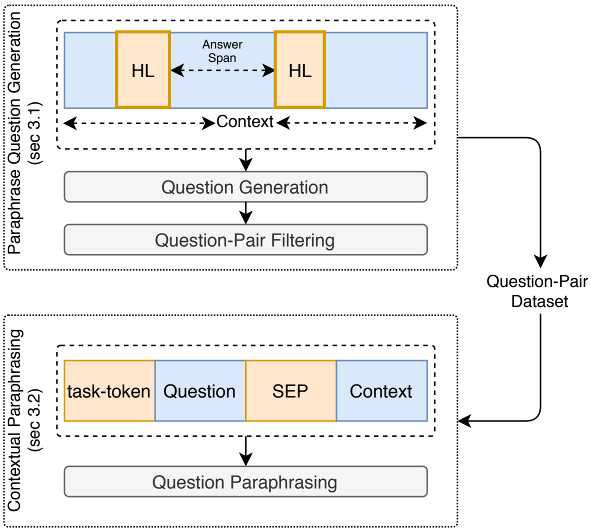

Figure 1: *ConQuest* framework. *ConQuest* has two main modules: 1) Paraphrase Question Generation to create the question-pair dataset and 2) Contextual Paraphrasing to expand and compress the questions.

3.1 Paraphrase Question Generation

The main objective of the Paraphrase Question Generation module is to create a dataset consisting of question-pairs, which can be used for downstream question paraphrasing tasks. Creating such a dataset manually is costly and difficult. To address this challenge, we introduce a novel approach based on answer-aware question generation. As shown in Figure 1, we first generate question-pairs of various lengths based on a reference passage and answer, and then filter by length to produce the final dataset. Specifically, to prepare the inputs and ensure that the question generation attends to the answer, we use a highlight token, **[HL]**, to surround the answer span in the input passage, following Klein and Nabi (2019). Then, we train the question generation model using the cross-entropy loss between generated questions and ground truth questions. Finally, a set of N questions are gener-

Model	Dispersity	Ent-1	Ent-2	Ent-3	Consistency	Fluency
QG	**17.4**	**1.04**	**1.14**	**1.15**	**93.2**	**82.5**
NMT	7.56	0.98	1.09	1.12	88.7	76.6
W2V	0.00	0.95	1.01	1.03	89.1	77.0

Table 1: Paraphrase question generation evaluation results.

ated using beam search.

The question-pair filtering step improves the generated question quality and length by 1) ranking the N generated questions by sequence length, and removing questions with less than 25% non-stop word token overlap with the shortest generated question to reduce the presence of paraphrases that are answer-inconsistent or semantically unrelated, and 2) removing questions less than three tokens longer than the shortest generated question to ensure length diversity. From the remaining questions, we create paraphrase sets between the shortest and longest generated questions. For cases with multiple short questions of identical length, we sample each individually to serve as references for the above filtering.

3.2 Contextual Paraphrasing

In this module, we train a contextualized multi-tag generative model for question paraphrasing (including question compression and expansion). By considering the context, the model has a better understanding of the important information which should be removed or added, and this, in turn, improves the overall quality of the paraphrased questions. To train this model, the dataset generated in Section 3.1 is divided into two subsets: 1) an expansion subset where the inputs are formed as **<EXPAND-TOKEN>Q<SEP>C** and the output is the longer version of the question and 2) a compression subset where the inputs are formed as **<SHORTEN-TOKEN>Q<SEP>C** and the output is the shorter version of the question. By using the **<EXPAND-TOKEN>** and **<SHORTEN-TOKEN>**, we provide a simple and effective method to disentangle the representations of these two tasks. Further, by using the multi-tag construct (i.e. training a single model on both tasks), we leverage the shared features of these tasks. Finally, a sequence-to-sequence BART model (Lewis et al., 2019) is trained using the cross-entropy loss between the generated questions and the targets. We evaluate our model and compare to alternative models in Section 4.2.

4 *ConQuest* Evaluation

To understand the performance of *ConQuest*, we have conducted a set of experiments using SQuAD v1.1, a commonly-used English question-answering dataset. In this section, we discuss the evaluation results for Paraphrase Question Generation and Contextual Paraphrasing.

4.1 Paraphrase Question Generation Evaluation

In this section, we evaluate the quality of our answer-aware QG-based paraphrase question generation model, relative to previous methods. The final synthetic paraphrase dataset contains 127,802 question-pair examples, created from 69,833 unique SQuAD v1.1 answer-context pairs, with an average compression rate of 63% between each generated pair.

We consider a number of automatic and human metrics to assess 1) diversity, 2) grammar and naturalness, and 3) consistency with the original SQuAD answer and context. The automatic metrics include Dispersity to measure the length distribution within generated paraphrase sets, and Ent-k (Zhang et al., 2018), an entropy-based diversity measure. Further, we consider Consistency and Fluency as human evaluation metrics. We compare our QG-based generation model to two previous approaches: back-translation via neural machine translation (NMT) (Xie et al., 2019), and synonym replacement using constrained word vectors (W2V) (Mrksic et al., 2016). More details about the evaluation metrics and implementation are presented in Appendices A and B.

The evaluation results in Table 1 show that answer-aware QG outperforms previous methods across all metrics considered. Specifically, our model achieves higher Dispersity, indicating, on average, a broader distribution of question lengths in each generated set. W2V, by comparison scores 0.0, as synonym replacement does not alter overall number of generated tokens. Our model also achieves higher Ent-1, Ent-2 and Ent-3 scores, indicating a lower degree of uniformity in generated

Compression Results	B1	B2	B3	B4	M	RL
BART-base (one-tags)	63.7	55.0	47.2	39.3	65.6	71.9
BART-base (two-tags)	**64.8**	**56.2**	**48.4**	**40.5**	**66.6**	72.3
T5-base (two-tags)	64.3	**56.2**	**48.6**	**40.6**	**66.6**	**72.8**
BART-base (two-tags, nocontext)	60.5	51.0	42.9	34.7	62.2	69.2
Expansion Results	B1	B2	B3	B4	M	RL
BART-base (one-tag)	62.2	53.9	46.8	40.3	65.4	68.4
BART-base (two-tags)	**62.8**	**54.3**	**47.1**	**40.6**	**65.8**	**68.7**
T5-base (two-tags)	61.9	53.4	46.2	39.4	66.1	68.5
BART-base (two-tags, nocontext)	51.6	41.2	32.7	26	53.4	57.0

Table 2: Contextual paraphrase generation evaluation.

n-gram distributions. While NMT scores higher than W2V, the results indicate that NMT is still limited to the semantic information contained in the original question, and cannot generate exceedingly diverse variants without considering the source context. Further, our model scores higher on both Consistency and Fluency, confirming that the QG-based questions are more grammatically correct, natural and answerable. Because the QG model is conditioned on the source answer and context, it retains consistency to a higher degree than approaches that are unaware of the answer and context.

4.2 Contextual Paraphrasing Evaluation

In this section, we discuss the performance of the Contextual Paraphrasing model on question compression and expansion tasks. Table 2 shows the results for this analysis. We consider multiple automatic metrics to assess the performance of the contextual question paraphrasing model, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and Rough-L (Lin, 2004). Some samples generated by *ConQuest* are presented in Section C

4.2.1 Base Model Evaluation

We have evaluated the contextual paraphrasing performance with BART (Lewis et al., 2019) and T5 (Raffel et al., 2019) as the base model. Compared to T5-base, the BART-base model shows similar performance for compression (64.8 BLEU-1 vs 64.3) and slightly better performance on the expansion task (62.8 BLEU-1 vs 61.9). Even though these results are close, T5-base has 220 million parameters, which is significantly larger than 139 millions for BART-base. Based on these results, we have chosen the BART model as the backbone of *ConQuest*.

4.2.2 Context vs No Context

ConQuest provides more meaningful question compression and expansion by considering the context during reformulation. To evaluate this factor, we performed experiments on samples with and without context using the BART-base model. As shown in Table 2, the contextualized text compression outperforms the non-contextualized model across all automatic metrics (e.g., improving BLEU-1 by 4.3 points). For the expansion task, our approach significantly outperforms the non-contextualized approach across all automatic metrics (e.g., 10.2 points improvement in BLEU-1). These results show that, by considering the context, our approach generates higher quality questions.

4.2.3 Single-tag vs Multi-tag Model

As discussed in Section 3.2, *ConQuest* utilizes a multi-tag input format to leverage the shared properties between the expansion and shortening tasks. To evaluate this factor, we have performed experiments with single tags and two tags, and compare their performance in Table 2. These results show that the two-tag model outperforms the one-tag model across all automatic metrics on both compression and expansion tasks. For example, the two-tag model results in 64.8 and 62.8 BLEU-1 score, which is 0.9 and 0.6 score improvement over the one-tag model. These results verify our assumption that using a multi-tag model leverages the compression and expansion task similarities and hence improves the model performance.

5 Conclusion

In this paper, we introduce *ConQuest*, a framework for generating synthetic contextual question paraphrase data. *ConQuest* first employs an answer-aware question generation model to create a dataset of diverse question-pairs. Then, it trains a multi-tag contextualized question paraphrasing model, which is able to control the length of the paraphrased questions. We have extensively evaluated *ConQuest* using the SQuAD v1.1 dataset. The results show that contextualized question paraphrasing results in higher performance across various automatic metrics (e.g., 4.3 and 11.2 points improvement in BLEU-1 compared to non-contextualized baseline). Further, our answer-aware QG-based data generation model achieves greater diversity, naturalness and consistency than previous question paraphrase models. Finally, we have performed and presented various ablation studies.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2019. Comqa: A community-sourced dataset for complex factoid question answering with paraphrase clusters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 307–317. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Andrea Gesmundo, Neil Houlsby, Wojciech Gajewski, and Wei Wang. 2017. Ask the right questions: Active question reformulation with reinforcement learning. *CoRR*, abs/1705.07830.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5917–5923. Association for Computational Linguistics.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. ChainCQG: Flow-aware conversational question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2061–2070, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *CoRR*, abs/2007.01282.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 6602–6609. AAAI Press.

Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds. *CoRR*, abs/1911.02365.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bang Liu, Mingjun Zhou, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. 2019. Learning to generate questions by learning what not to generate. In *WWW '19: The World Wide Web Conference*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xiyao Ma, Qile Zhu, Yanlin Zhou, and Xiaolin Li. 2020. Improving question generation with sentence-level semantic matching and answer position inferring. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8464–8471. AAAI Press.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. *arXiv preprint arXiv:2104.06828*.

Sungrim (Riea) Moon and Jungwei Fan. 2020. How you ask matters: The effect of paraphrastic questions to BERT performance on a clinical SQuAD dataset. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 111–116, Online. Association for Computational Linguistics.

Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. Counter-fitting word vectors to linguistic constraints. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 142–148. The Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Stalin Varanasi, Saadullah Amin, and Guenter Neumann. 2020. CopyBERT: A unified approach to question generation with self-attention. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 25–31, Online. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. *CoRR*, abs/1908.11828.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *CoRR*, abs/1809.05972.

A Evaluation Metrics

We evaluate different parts of our approach on a number of automatic and human metrics to assess 1) diversity, 2) grammar and naturalness, and 3) consistency with the target answer and context. Specifically, to evaluate the paraphrase question generation, we use Dispersity and Ent-k metric for automatic scores and Consistency and Fluency for human evaluation (used in Section 4.1). To evaluate the quality of the contextual paraphrase generation, we use BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and Rough-L (Lin, 2004) (used in Section 4.2).

The dispersity is defined as,

$$dispersity = \frac{std(q_l)}{mean(q_l)} * 100 \quad (1)$$

where q_l is the set of generated question lengths. The Ent-k metric, describing diversity of generated questions via n-gram frequencies (Zhang et al., 2018). To calculate these metrics, we generate 8 reference paraphrase questions for each of 2000 examples subsampled from SQuAD v1.1. Each method is given a common set of 2000 source questions; In addition, the QG-based generation method is only given the corresponding answer and context passages for each question, and not shown the original SQuAD questions themselves.

Human evaluation assesses the models' abilities to capture semantic meaning within paraphrases while retaining grammatical accuracy. Specifically, we use Consistency, and Fluency to measure the quality of the generated questions in relation to the answer and context. We have used Mechanical Turk for this evaluation. For each task and each generation model, we randomly selected 200 generated samples, with each sample scored by ten different Amazon Mechanical Turk (Turkers). For Consistency, Turkers decide a generated question is answerable or not based on the provided passage and answer. Based on choice between "Consistent" (score 3), "Can not decide" (score 1), and "Inconsistent" (score 0), we get the Consistency score by divided the total score with the theoretically possible maximum score (3*total question number). For fluency, the Turkers evaluate how many errors in grammar level, lexical leval, semantic leval combined, given the passage. Based on choice between "No errors" (score 3), "one error" (score 1), and "more than one error" (score 0), similar to the Consistancy calculation, we have the final Fluency score by divided the total score with the possible maximum score (3*total question number).

B Experimental Details

In this section, we provide additional experimental details which is of use in replicating the evaluation results.

Paraphrase Question Generation: We consider T5-base (220M params) for question generation. The model is trained for 5 epochs. We apply AdamW optimizer (Loshchilov and Hutter, 2019), and the number of warmup ratio is set to be 0.1. Learning rate is tuned between 2e-5 5e-5. The dropout ratio is set to be 0.1. We test and chose the better decoding method for the model. Specifically, we decode questions via beam search of size 4. We encode the model input as: **CONTEXT[SEP]ANSWER**, and use a cross entropy loss between generated tokens and tokens in ground truth question for each example in the inverted SQuAD training set.

Paraphrase Question Generation Baselines: For our dataset evaluation, we consider two previous paraphrase generation approaches. The first is back-translation via neural machine translation (NMT), in which a question, q, written in English is translated to French, and then back to English to generate a syntactic variant, q', with the same semantic meaning. We use the UDA package (Xie et al., 2019), with WMT'14 English-French translation model checkpoints. For decoding, we use random sampling, with a sampling temperature of 0.8. The second approach we consider is synonym replacement using constrained word vectors (W2V), trained specifically for the synonym replacement task (Mrksic et al., 2016). We restrict synonym replacements to non-stop words, and target a 25% non-stop word replacement rate for each question. These results are discussed along with the proposed QG method in following sections.

Contextual Paraphrasing: To train the contextual paraphrasing model, we have explored BART-base (140M params) (Lewis et al., 2019) and T5-base (220M params) (Raffel et al., 2019), as described in Section 4.2. Each model is trained for 6 epochs. AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate of 1e-4 (with a linear scheduler), β_1 of 0.9, and β_2 of 0.999. To generate the samples, we use nucleus sampling (Holtzman et al., 2020) with top-p as 0.92. These hyperparameters are chosen empirically. On average training

Table 3: Samples generated by the model

Expansion Samples
Before: Who excavated ancient Ur? **After:** Who excavated ancient Ur and published The Art of the Middle East, including Persia, Mesopotamia and Palestine?
Before: Which system lacks a distinctive future tense? **After:** Which system lacks a distinctive future tense (the present tense serves here) and features special forms to express an action performed by an undetermined subject (the "impersonal"?
Before: Who encouraged Tony de Brum? **After:** Who encouraged Tony de Brum to turn the crises into an opportunity to promote action against climate change?
Compression Samples
Before: Who coined the term Hellenistic to refer to and define the period when Greek culture spread in the non-Greek world after Alexander's conquest? **After:** Who coined the term Hellenistic?
Before: The Alps are the highest and most extensive mountain range system that lies entirely where? **After:** Where do the Alps lie?
Before: What is far superior to classical thermodynamics in that glass breaking behavior can be explained by the fundamental laws of physics paired with a statistical postulate? **After:** What is far superior to classical thermodynamics?

for 6 epochs has taken 6 and half hours in Amazon EC2 p3.2xlarge instances with Tesla V100 GPUs. The data used for this model is divided in 90% training and 10% test data. Further, 5% of the training data is considered as validation set. This split will be released upon acceptance of the paper.

C Generated Samples:

In this section, we provide some samples generated by the contextual paraphrasing model (shown in Table 3). As the samples show, in the expansion task, the model uses some additional info from the context to expand the questions. For the compression task, *ConQuest* removes some of the information provided in the question; however tries to keep the most important part of the information.

D Ethics/Broader Impact Statement

The results presented here demonstrate the efficacy of *ConQuest* in producing challenging datasets by leveraging a question paraphrasing model. This model enables higher robustness of Question Answering models in real-life applications. As part of our future work, we plan to 1) extend this framework to denoising and noisifying tasks, 2) extend to tasks other than single-turn QA(e.g., conversational QA and text summarization), and 3) end-to-end and joint training of QA and QP models. With respect to ethical considerations, our framework has the same considerations as a general QA and QG model and hence, we inherit any ethical shortcomings that exists in the source dataset.