Evaluating Question generation models using QA systems and Semantic Textual Similarity

Safwan Shaheer
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Dhaka, Bangladesh
safwan.shaheer@g.bracu.ac.bd

Ishmam Hossain
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Dhaka, Bangladesh
ishmam.hossain@g.bracu.ac.bd

Sudipta Nandi Sarna
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Dhaka, Bangladesh
sudipta.nandi.sarna@g.bracu.ac.bd

***Abstract*—Question generation based on conversational context is a difficult problem to solve. A widely used technique for generating quality questions using fine-tuned models rely on a suitable answer and the context, usually the passage. But when it comes to conversational setting the questions generated is not of the highest quality as it lacks the contextual element in the question, especially due to the lack of co-reference resolution of the entity. Furthermore in most of the evaluation techniques for generating questions there seems to be a lack of utilizing powerful question answering system to judge the answerability of the question generated. The most prevalent metric used for judging machine generated text against human golden standard, BLUE unfortunately doesn't factor in whether a question answering system would be able to answer the question, but mostly on the number of substrings that matches against each other. Various question generation models following a generalized encoder-decoder architecture was evaluated using semantic textual similarity on both the gnerated questions and the generated answer, Although higher parameters of a model usually lends to better performance our experiment displayed that such is not always the case atleast when there is massive amount of context absent.**

***Index Terms*—Question Generation, Semantic Textual Similarity, Question Answering**

I. Introduction

Automatic Question Generation system aims to generate a valid, fluent and intelligible question, with minimal human intervention using mostly machine learning models. according to a given passage and in most of the case an additional context using the target answer. The target answer is required as otherwise the model would lack the context required when generating questions. Such system is able to generate question provided the passage and answer as context. In recent times we have seen a massive explosion of open source models which are hosted on the popular service Huggingface that are responsible for generating text by providing some context information or via prompting. Recently Transformers have demonstrated to be trainable faster (due to pretaining on massive corpus) and arguably better than traditional RNN architectures. For example Grover et al. [1] demonstrated such capabilities of such model by fine-tune the pre-trained T5 model for a downstream job of question generating, which works very well even on unobserved data and creates well-structured, linguistically accurate queries. In most of the cases such models are fine-tuned version of an already pretrained model, using the popular Stanford SQuAD dataset for training. We wanted to judge the capabilities of such system against a lot more unpredictable and less contextual dataset for example the likes of CoQA. The ability to generate questions is useful in a variety of contexts, including autonomous teaching systems, enhancing the functionality of question-answering models, and providing chatbots with the ability to take the lead in a discussion. It can be used as a major tool for automating the tedious process of generating high-quality academics questions paper.

II. Research Objectives

In this research we want to evaluate Transformers-based Question generation model using a different approach. We want to make use of Question Answering model to extract answers for the generated questions. The semantic textual similarity between the question and generated question is compared using cosine-similarity after generating semantically correct sentence embeddings using a SBERT, a modification of the pre-trained BERT network (Reimers et al.) [2]. The goal is to explore the possible quality gap between the generated questions given a contextual and non-contextual setting where there are a lot of ambiguity starting from basic co-reference resolution. A dataset like CoQA demonstrate that conversational questions include difficult phenomena that are absent from available reading comprehension datasets, such as coreference and pragmatic reasoning [3]. Furthermore also suggest some possible area of prospect solutions that might make the question generation part more robust using co-reference resolution and injecting root word into the prompt

using dependency parsing, basically reducing ambuigity and providing more context.

III. Problem Statement

Co-reference resolution is the capacity to recognize and comprehend the connections between things stated in a text. In the context of question creation, this may be a difficulty for the models used, since they may not always be able to effectively determine which elements in a text are connected to one another, resulting in the development of questions that are not properly aligned with the text's meaning. This may lead to question generating systems that generate vague or confusing questions. BLUE is a popular metric used for automatically evaluating text generation using a language-independent, fast and fairly cheap way that has high similarity with how humans evaluate text [4]. But unfortunately BLEU metric is not effective at capturing the nuances and complexities of natural language, so it may not accurately reflect the quality of the generated text. According to a study by Sulem et al. found that BLEU often penalizes sentences that are simpler due to negative correlation with simplicity even though they contain the same semantic meaning [5]. Even after filtering out extremely ambigious questions from the conversational dataset like CoQA the current question generation models are fairly ineffective when generating questions. This is mostly due to the dependencies of the previous conversation that was done so far, as well as the volume of entities referenced previously using coreferences which is usual how humans converse [6]

IV. Related Works

ROUGE is a BLEU variant that puts a greater focus on recall than accuracy. This implies that it looks at how many n-grams from the reference translation appear in the output rather than the contrary. [7]. Novikova et al. (2017) demonstrate that BLEU, in addition to a few other regularly used metrics like ROGUE, does not translate very well to human assessments when it comes to assessing natural language generation (NLG) activities [8]. Rathod et al. proposed automatic metrics grounded in desirable properties of the generated questions, answerability, where the generated question is fed to a QA model to produce the correct answer for each question since both the generated question and the actual question are intended to have the same answer [9]. Past work by Yuan et al demonstrated the significance of evaluating the accuracy of qg models using qa models to ensure the generated questions are answerable [10]. Semantic textual similarity (STS) aims to detect similarity in meaning between two texts. A rigorous case study by Marie Stephen shown that SBERT on average the best STS metric available [11]

V. Dataset

We have evaluated the open source hosted models on the following dataset (except SQuAD). The following subsection gives a brief overview of each of them.

A. SQuAD

The SQuAD is a large dataset of over 100,000 crowd-sourced question-answer pairs, based on a set of Wikipedia articles. (Rajpurkar et al. 2016) [12]. For numerous reasons, SQuAD (Stanford Question Answering Dataset) has become the go-to dataset for fine-tuning question generation models. One of the primary reasons is that SQuAD has a vast quantity of high-quality training data on a variety of themes and difficulty levels. This makes it ideal for training question-generation models that can handle a broad variety of inputs while producing correct replies. Furthermore, SQuAD is well-known and frequently utilized in the natural language processing field, so many researchers are acquainted with it and have used it to assess their models. As a result, it is a useful and generally acknowledged dataset for fine-tuning question generation algorithms. In the figure 1 a sample from the dataset is shown:-

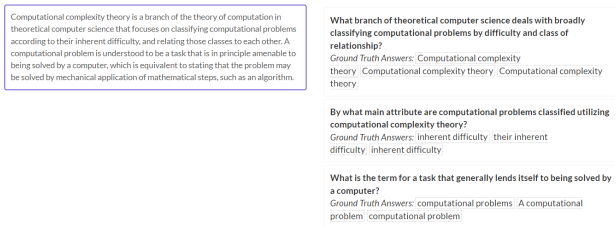

Fig. 1: Squad Dataset Preview

B. CoQA

CoQA is a massive dataset used to train Conversational Question Answering systems. It comprises over 127,000 questions and responses from over 8000 interactions. CoQA has several complex phenomena that are not seen in previous reading comprehension datasets, such as coreference and pragmatic reasoning. Each response is accompanied by an evidence subsequence marked in the paragraph (Reddy et al.) [3]. In the figure 2 a sample from the dataset is shown:-

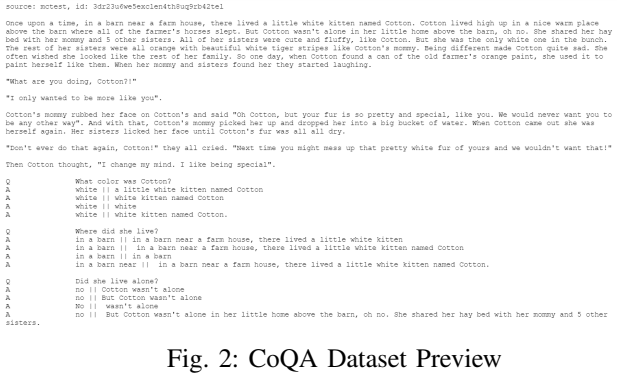

Fig. 2: CoQA Dataset Preview

C. MLQA

The MLQA provides sample responses to frequently requested questions in English, Arabic, German, Spanish, Hindi,

Vietnamese, and Simplified Chinese. It is made up of about 12k English instances, with each QA instance supporting an average of four separate languages at the same time. Cross-lingual MLQA supplements extractive QA datasets. (Lewis et al.) [13]. In the figure 3 a sample from the dataset is shown:-

Fig. 3: MLQA Dataset Preview

D. Pre-processing

Both MLQA and CoQA dataset had their own schema thus in order to feed them to QG models we would need to pro-process then and convert them to a unified and standard format of question, answer and context tuple. Since CoQA dataset is conversational in nature, there might be context missing in the actual question since it might reference entities described or explained in the preceding questions/answers. Furthermore questions which might be contextually correct will not be enough for the QA models for example "Why?" & "How?" the likes of which are found plenty in the dataset. Thus we have filtered the short-questions (indicating that they lack the appropriate context) and checked whether the question began with natural question words (What, when, how, where, why) We have also excluded the answers which is deemed as unknown. Further more since its a conversational dataset and there is natural flow, its common to encounter pronouns. We want to keep such references as its crucial for us to understand and evaluate the performance of the QG models when they are faced with ambiguity. We also skip the answers that are over a certain threshold of words. As for the MLQA questions since its cross-lingual in nature we have only selected the en-en subset of the dataset, which contains both question and answer alongside the corresponding passage in english language.

VI. Methodology

We will briefly describe the various transformer and encoder-decoder based QG models that were used while performing the experiment. The implementation and usage of such models are publicly available on our github repository and corresponding colab notebooks.

A. T5-base finetuned for QG

The T5 transformer model is a large-scale natural language processing model developed by Google in 2020. It is an extension of the popular Transformer model architecture, which uses self-attention mechanisms to process input text. T5 is capable of wide range of NLP tasks including language translation, text summarization, and question answering, using a single, unified model architecture. This is a significant advancement over previous natural language processing models, which often required task-specific architectures and training regimes. T5's success is due in part to its enormous size and the use of advanced techniques such as Transfer Learning, which allows it to effectively learn and adapt to new tasks using a pre-trained model as a starting point. Attaching a prefix to the input for each activity, such as translation:, enables T5 to perform admirably right out of the box on a wide variety of tasks. Summarizing: summarize:.... . [14]. A fine-tuned version of Google's T5-base fine-tuned is able to generate questions providing the answer and the passage [15]

B. BART-base finetuned for QG

BART excels in text production (summarization, translation) and comprehension (e.g. text classification, question answering) [16]. We have used a fine-tuned model of BART hosted on hugging face that given a section of a passage (context), the model is tasked to generate questions from the passage about the selected section or context [17]

C. GPT2 finetuned for QG

GPT-2 is a transformers model that was self-supervisedly trained on a massive corpus raw texts, without any human labeling, with an automated mechanism generating inputs and labels from those texts [18]. In order to produce question-answer pairs, Krishna and Iyyer (2019) designed a pipeline, where the input text is converted into a question using GPT-2, and BERT provides the output answer [19]. We will make use of a model hosted on huggingface that uses GPT2-small (124M parameters) for question generation given the context sentence and the answer word [20].

D. distilbart-cnn finetuned for QG

This model is a condensed version of the BART-large-cnn model that has been pre-trained on the English language and fine-tuned on CNN Daily Mail. A fine-tuning procedure was performed on the SQuAD dataset by utilizing the sshleifer/distilbart-cnn-6-6 summarization checkpoint. This process resulted in the production of the distilbart-qgen-6-6 model checkpoint [21]

E. Proposed Methodology

We will randomly sample 1000 items from each dataset (CoQA and MLQA) and feed them to the Question generation models. The generated questions is then matched against the actual question using sentence transformer's semantic textual similarity mechanism. Semantic Textual Similarity (STS) assigns a score on the similarity of two texts. A normalized score within the range of 0-1 is returned. The generate questions will also be passed to a question answering model, and the answer will be matched against the actual answer using the same technique. Before calculating the STS value both the text would be normalized using the standard SQuAD normalization script provided for evaluation, including lowercasing, removing duplicate stopwords and so on. Furthermore we will visualize a barchart to understand whether the nature of the questions dictate the answerability of the questions generated and the relation between using a two different sizes of the

same transformer model. We have decided to skip standard precisipn driven metrics like BLUE and recall driven metric ROGUE based on the above disucssions. In the figure 4 we have shown our proposed methodology

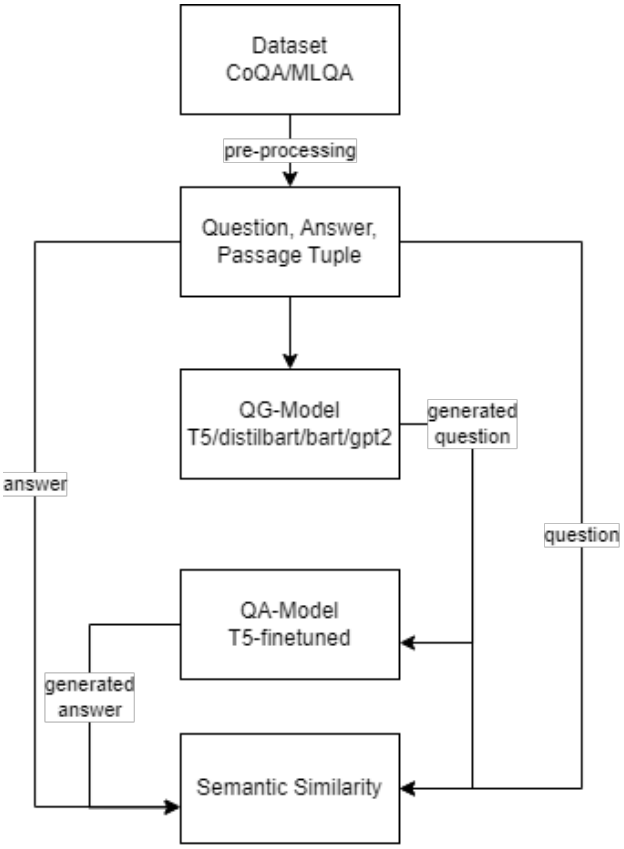

Fig. 4: Methodology

VII. Experimental Result Analysis

model	mlqa_avg_qsm	mlqa_avg_asm	coqa_avg_qsm	coqa_avg_asm
bart-base	0.597067	0.850951	0.388793	0.654058
distilbart	0.613098	0.838836	0.368486	0.644957
gpt2-small	0.474006	0.663012	0.392500	0.625554
t5-base	0.635866	0.861155	0.435214	0.692511
t5-large	0.601336	0.836608	0.434320	0.735961

QG Model average semantic similarity score across datasets

The above table depicts the average similarity score across the two dataset based on two metrics. Each row corresponds to the qg model used and the value of each cell is a mean value normalized within the range 0-1. Columns prepended with ‘mlqa‘ denotes the MLQA dataset where ‘coqa‘ denotes the CoQA dataset. qsm is used to denote ‘question semantic similarity‘ between the generated question from the QG model and the actual question, while ‘asm‘ is used to denote the same for the answers. Almost in all cases the models had a tougher time generating questions from the coqa dataset which was as expected due to its low context nature. Interestingly there was no noticeable difference between the ‘t5-base‘ and ‘t5-large‘ model, even though the latter is 3x times larger, and in some cases the lower parameter model slightly outperforms.

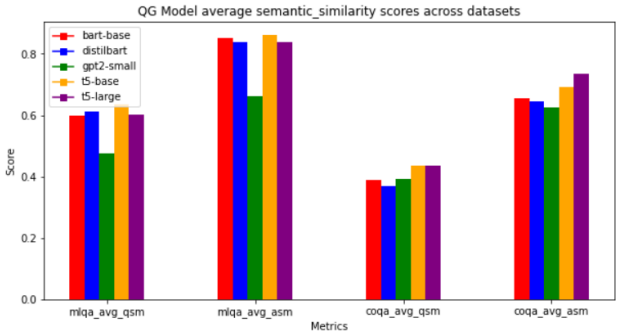

Fig. 5: QG Model average semantic similarity score across datasets

VIII. Conculsion

We demonstrated that generating questions from a fine tuned model can be quite challenging from a dataset like CoQA, since its conversational in nature. Even for humans this will prove to be difficult as the questions most of the time lack the necessary context to attach it with the answer. We performed a more automated approach of scoring the questions generated than relying on matching the actual ngram using PINC, BLUE or ROGUE metrics. We demonstrated that due to the lack of context question generation models have a hard time dealing with CoQA dataset. Several other popular datasets for example Natural Questions which relies on open domain question answering [22] and such could be utilized during the collection of randomized samples when generating questions. A larger volume of sample data could be used for generating the evaluation report for the models which was skipped due to hardware constraints. T5-B11 is the checkpoint of T5 transformer model with 11 Billion parameters, which has acheived SOTA result on question answering. Such resource extensive model was also skipped due to its gargantuan size but can also be utilized in the question answering step. A larger and high-quality neural question generation model like ProphetNet which was pretained on a base and large scale dataset of 160GB that achieves SOTA result on question generation tasks [23], can also be utilized. Further studies can be conducted using the same approaches mentioned above on how current models performs by modelling long documents that can find relevant context for generating the question from the provided passage [24].

References

[1] K. Grover, K. Kaur, K. Tiwari, Rupali, and P. Kumar, “Deep learning based question generation using t5 transformer,” in *Advanced Computing*, D. Garg, K. Wong, J. Sarangapani, and S. K. Gupta, Eds. Singapore: Springer Singapore, 2021, pp. 243–255.

[2] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *CoRR*, vol. abs/1908.10084, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[3] S. Reddy, D. Chen, and C. D. Manning, “Coqa: A conversational question answering challenge,” *CoRR*, vol. abs/1808.07042, 2018. [Online]. Available: http://arxiv.org/abs/1808.07042

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[5] E. Sulem, O. Abend, and A. Rappoport, "BLEU is not suitable for the evaluation of text simplification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 738–744. [Online]. Available: https://aclanthology.org/D18-1081

[6] Y. Gao, P. Li, I. King, and M. R. Lyu, "Interconnected question generation with coreference alignment and conversation flow modeling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4853–4862. [Online]. Available: https://aclanthology.org/P19-1480

[7] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 150–157. [Online]. Available: https://aclanthology.org/N03-1020

[8] J. Novikova, O. Dušek, A. Cercas Curry, and V. Rieser, "Why we need new evaluation metrics for NLG," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2241–2252. [Online]. Available: https://aclanthology.org/D17-1238

[9] M. Rathod, T. Tu, and K. Stasaski, "Educational multi-question generation for reading comprehension," in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 216–223. [Online]. Available: https://aclanthology.org/2022.bea-1.26

[10] X. Yuan, T. Wang, C. Gulcehre, A. Sordoni, P. Bachman, S. Zhang, S. Subramanian, and A. Trischler, "Machine comprehension by text-to-text neural question generation," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 15–25. [Online]. Available: https://aclanthology.org/W17-2603

[11] M. S. Leo, "Semantic textual similarity," https://towardsdatascience.com/semantic-textual-similarity-83b3ca4a840e, Apr. 2022, accessed: 2022-12-10.

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," *CoRR*, vol. abs/1606.05250, 2016. [Online]. Available: http://arxiv.org/abs/1606.05250

[13] P. S. H. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: evaluating cross-lingual extractive question answering," *CoRR*, vol. abs/1910.07475, 2019. [Online]. Available: http://arxiv.org/abs/1910.07475

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[15] M. Romero, "T5 (base) fine-tuned on squad for qg via ap," https://huggingface.co/mrm8488/t5-base-finetuned-question-generation-ap, 2021.

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461

[17] kaejo98, "Bart (base) fine-tuned on different qa dataset," https://huggingface.co/kaejo98/bart-base_question_generation, 2022.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[19] K. Krishna and M. Iyyer, "Generating question-answer hierarchies," *CoRR*, vol. abs/1906.02622, 2019. [Online]. Available: http://arxiv.org/abs/1906.02622

[20] D. Khashabi, "Gpt2 (small) fine-tuned on qa dataset," https://huggingface.co/danyaljj/gpt2_question_generation_given_paragraph_answer, 2021.

[21] G. Singh, "distilbart-cnn fine-tuned on squad dataset," https://huggingface.co/gpssohi/distilbart-qgen-6-6, 2021.

[22] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, 2019. [Online]. Available: https://aclanthology.org/Q19-1026

[23] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2401–2410. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.217

[24] L. A. Tuan, D. J. Shah, and R. Barzilay, "Capturing greater context for question generation," *CoRR*, vol. abs/1910.10274, 2019. [Online]. Available: http://arxiv.org/abs/1910.10274