2024 27th International Conference on Computer and Information Technology (ICCIT)
20-22 December 2024, Cox's Bazar, Bangladesh

Optimizing Resource-Scarce Language Question Generation on Bengali : A Comparative Study of Transformer-Based Models

Fahad Siddique Faisal
Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
Fahadsid1770@gmail.com

Syed Nakibul Islam
Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
s.nakib30@gmail.com

Md. Farhad Hossain
Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
farhad.hossain@cuet.ac.bd

Dr. Md. Azad Hossain
Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
azad@cuet.ac.bd

Abstract—The process of creating potential questions from text or other types of data, like documents, images, graphs, etc., is commonly referred to as Automatic Question Generation (AGQ) or simply Question Generation System. The objective of this research is to explore the application of Question Generation (QG) within the frequently asked questions (FAQ) domain, ultimately leading to the design and implementation of an automatic FAQ generation system. Question Generation studies in Finnish, Swedish, Turkish, and English face limitations due to poor machine-translated datasets, inadequate evaluation metrics, low text quality, limited results, and restricted tabular data use during COVID-19. Eliminating these limitations, we have considered this and employed a series of refined text-to-text encoder-only and encoder-decoder transformer-based models named IndicBart, SahajBart, Biswabangla, BanglaBERT, BanglaT5, and lastly Google's mT5-base model. The proposed bn-TOK + mT5-base + Beam Search model scores Rouge 37.225 and SacreBLEU 14.202 with METEOR 0.216 and demonstrates superior performance in given metrics, balancing lexical accuracy, structural coherence, and semantic meaningfulness when the context of the text and the answers are given to create questions that seem humane.

Index Terms—Automatic Question Generation (AGQ), Transformer-Based Models, Low-Resource Languages, Bangla Question Generation

I. Introduction

Question generation (QG) is a crucial process in natural language processing, particularly in Bengali. It involves creating questions based on a specific passage or context, which is increasingly popular in both academic and business sectors. Early adopters focused on creating closed-ended or gap-filling questions, but current research faces several challenges.

Bengali presents significant NLP challenges due to its complex verbal morphology, where a single root generates hundreds of forms, and compound verbs create new meanings. Its flexible Subject-Object-Verb (SOV) word order and heavy reliance on context, including implicit pronouns, complicate meaning resolution and question generation.

The intricate Bengali script, with visually similar characters, and its nuanced use of particles and modifiers further increase complexity. Bengali interrogatives, like "what" and "why," demands attention to grammar, pragmatics, and formality. Limited resources, such as annotated datasets, parsing tools, and tailored pre-trained models, highlight the need for improved tokenization and specialized solutions.

This study aims to tackle significant challenges by introducing advanced multilingual and Bengali transformer models designed for generating context-sensitive, answer-oriented questions. The objectives are threefold: (1) to improve question-answering systems by addressing the limitations of existing methods, laying the groundwork for more accurate and robust solutions, (2) to enhance Bengali chatbot performance, enabling them to handle complex queries more effectively, and (3) to facilitate data augmentation for creating enriched resources such as articles and textbooks, thereby contributing to advancements in education and research.

The objectives of this research include employing both encoder and decoder architecture instead of encoder-only transformer, implementing bn-tokenizer for better Bangla tokenization, and specifying the answer via masking in the dataset to give it answer-aware status.The main contributions of this work are:

- Employing both encoder and decoder architecture instead of encoder-only transformer for better contextual understanding.
- Implementing bn-tokenization instead of T5 tokenizer and Bert-tokenizer for better Bangla Tok-

enization.
- Implementation of the model by specifying the answer via the masking process in the dataset gives it the answer-aware status.

In this paper, Section II reviews related work and previous research on this topic. Section III details the methodology, including data processing, dataset description, and model implementation. Section IV presents the experimental setup and results. Finally, Section V offers a conclusion based on the findings.

II. Related Works

In 2010, Rus et al. introduced "The First Question Generation Shared Task Evaluation Challenge" [1], pioneering question generation from reading passages using linguistic elements and templates. That same year, Heilman et al. developed a statistical ranking method for question generation [2], leveraging features like question type, word importance, and language model scores to identify optimal questions.

A. Generating Question In Non-English Languages

Numerous researchers have shown interest in the topic of answering questions, especially in languages other than English. To train question-answering systems, Mayeesha et al. translated SQUAD 2.0 into Bengali [3], and Bhattacharjee et al. benchmarked the BanglaBERT model [4]. On the other hand, the problem of question generation has not been tackled. Question generation in other languages has been attempted: factual question generation [5], factoid-based question generation [6], distractor generation [7], transformer-based question generation in Turkish [8], and cross-lingual and code-mixed situations [9]. Answer-aware QG tasks have been implemented in English utilizing many iterations of the T5 model. To obtain the most accurate findings for non-English languages, T5 and other models have been used. Factoid question generation in Finnish has been achieved by applying BERT and GPT-2-based models, with the highest Rouge-L value of 0.33 [6].

B. Bangla Natural Language Generation

Natural language generation (NLG) is seeing a spike in attention due to the popularity of transformer models like OpenAI's GPT-3 [10]. NLG systems produce natural language text for use in writing, storytelling, screenplays for movies, poetry, dialogue, and answering questions to achieve communication objectives [11]. The well-known AI support tools are Github's Copilot [12] and OpenAI's GPT-3. These systems can alter the persona, tone, topic, and content of created text by using controlled text generation. However, NLG systems become more complicated due to issues such as hallucinations [13], choosing the right evaluation criteria [11], [14], and erroneous or non-factual information. Large-scale datasets for machine translation and abstractive text summarization [15] have been made available, but there is still much to learn about the wide-ranging use of transformer models for tasks like text generation, question or conversation creation, and text conversion. The "question generation" (QG) challenge, which attempts to automatically produce questions based on a provided background paragraph, is the main subject of this work. While the IndicNLG benchmark [16] has investigated Bengali question production among other tasks, the Bangla NLG benchmark [17] covers tasks like machine translation, question answering, dialogue generation, and cross-lingual summarizing.

Previous Question Generation (QG) studies across languages have faced significant limitations. Finnish QG struggled with poor machine-translated datasets and inadequate evaluation metrics like BLEU. Swedish QG was hindered by low text quality and reliance on an encoder-only model. Turkish QG lacked diversity in question generation, while English QG during COVID-19 tasks suffered from limited use of tabular data, reducing question complexity. These challenges highlight issues in dataset quality, model design, and evaluation methods.

For Bangla, previous studies relied on pre-trained models like BanglaT5 and mT5 but neglected essential preprocessing techniques such as stemming, lemmatization, tokenization, and advanced decoding. Our work overcomes these gaps by integrating advanced preprocessing tools and powerful decoding algorithms to generate contextually relevant and grammatically accurate questions.

III. Methodology

The study leverages the T5 model, a pre-trained text-to-text transformer, to generate questions based on context and answer cues. Known for surpassing existing transformer models in text production tasks [18], T5 was fine-tuned for Bengali question generation using a dataset adapted from a question-answering corpus. The model was pre-trained on a vast multilingual dataset, including Bengali, by predicting masked words from surrounding context, optimizing its language modeling capabilities. In the experiment, the fine-tuned T5 model was provided with input consisting of context and answers to generate questions with outputs through automated metrics.

A. Selecting data and dataset

No dataset exists for Bengali question generation, so the BQA dataset [19], derived from SQUAD 2.0 [20], was adapted for this task. Validation and test datasets were sourced from TidyQA's secondary gold passage task. With 132,777 samples, BQA aligns with research in Arabic, Hindi, and Finnish. BanglaT5 was used with this dataset in the BanglaNLG benchmark to evaluate Bengali question answering.

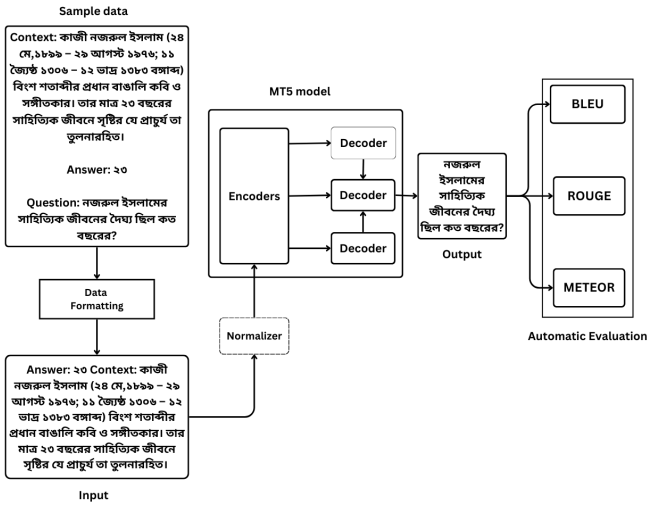

Fig. 1. Simplified Methodology for Entire QG system.

B. Preprocessing

Question generation research has been undertaken using two approaches: answer-agnostic and answer-aware. In the answer-agnostic [21] scenario where the answer is not considered, input passages are supplied without a provided answer. However, in this current scenario where the answer is considered, the answer to the question is already provided. This strategy is more straightforward to assess because it includes a reference question throughout the finetuning stage. Two ways are recommended for managing the many-to-one relation: All questions per line (AQPL) and One question per line (OQPL).

TABLE I
Comparison of Dataset-BQA and Final Dataset

Dataset	Train	Validation	Test
Dataset-BQA	127,771	2,502	2,502
Final Dataset	13,549	4,523	3,250

The trials are conducted within the framework of one specific question for each passage. Following the application of a filtering process, a total of 13,549 samples are available for training purposes, whereas 4,523 samples are allocated for validation. The dataset has been pre-processed to meet the necessary format for generating questions that are aware of the answers. Only acceptable response pairs and filtered passages with a maximum context length of 500 characters have been retained. Unanswered questions are also excluded.

The data sample is provided in its original form, with the "answer:" and "context:" sections separated. The "target" refers to a question sentence.

C. The Model Architecture

This work focuses on generating questions using a modified dataset comprising IndicBart, SahajBert, Biswabangla, BanglaBERT, and BanglaT5 [19]. Among

Input :
answer: মৌলভী আবদুস সামাদ
context: মতিউর রহমান ১৯৪১ সালের ২৯ অক্টোবর পুরান ঢাকার ১০৯, আগা সাদেক রোডের "মোবারক লজ"-এ জন্মগ্রহণ করেন। তার পিতার নাম মৌলভী আবদুস সামাদ এবং মাতার নাম সৈয়দা মোবারকুন্নেসা খাতুন।

Target:
মতিউর রহমানের বাবার নাম কি ?

Fig. 2. Sample input state before and after the data processing.

these, MT5, a multilingual version of the T5 transformer, is pre-trained on the mC4 dataset, which includes data from 101 languages sourced from Common Crawl. T5 itself is a text-to-text transfer transformer designed for English and addresses diverse NLP tasks by converting them into text-based problems. For instance, classification tasks are framed as text generation, assigning labels like "positive" or "negative," while QA tasks involve generating answers based on the given input text.

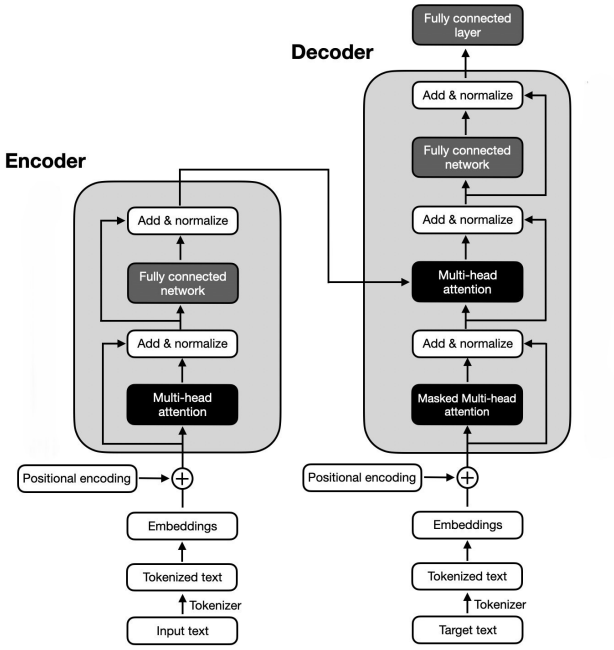

Fig. 3. Text-To-Text encoder-decoder Transformer architecture [22]

The Bengali question-generation process ensures accuracy and relevance through structured steps. It starts with data formatting, organizing input data (context and answer) into a coherent structure for processing. For example, a context about Kazi Nazrul Islam's biography and a related answer are formatted to prepare for normalization and subsequent steps.

1) Normalization: Following formatting, the data undergo normalization through the "bnUnicodeNormalizer" . This step is essential to standardize the text

by removing inconsistencies such as punctuation marks, irregular whitespace, and variations in the text casing. Normalization guarantees that the text is uniform, which is critical for accurate tokenization.

2) Tokenization : This is the next step, tokenization, involves breaking down the text into individual tokens using a tokenizer specifically designed for Bengali, bn-tokenization. This process converts the text into a sequence of tokens, facilitating easier processing by the model and enabling it to handle linguistic nuances specific to the Bengali language.

3) Positional Encoding: Positional encoding allows the model to understand the sequence of tokens, thereby enhancing its ability to generate contextually appropriate outputs.

4) Encoder : Muti-head self Attention Mechanism: The encoded tokens enter the encoder, where multi-head self-attention captures relationships across the sequence, and feed-forward networks refine representations for better contextual understanding.

5) Decoder : Masked-Attention Mechanism: The decoder ensures that each token generation step considers only the previous tokens, preserving the autoregressive nature of the task. The decoder-encoder multi-head self-attention integrates contextual information from the encoder's output, enhancing the relevance of the generated question.

D. Decoding Algorithms

Decoding algorithms generate questions by predicting the next word at each time step. Greedy decoding selects the most likely word at each step, while Beam Search [23] evaluates multiple options to find the hypothesis with the highest cumulative likelihood. Sampling methods like top-k [24] and top-p [25] randomly choose words from the most probable set, with top-k and top-p values typically set between 10–90 and 0.9–1, respectively. To avoid overly long outputs, the maximum length is capped at 50 tokens. Our research prioritizes Beam Search for its superior context-awareness and ability to select optimal options.

IV. Experiments

A. Implementation Settings

This experiment uses Google Collaborators, GitHub, Hugging Face, and WandB platforms for data analysis, machine learning, and instructional activities, xl-sum for Multilingual Summarizing, Unidic for segmentation, punkt for tokenization, and normalizer for data preparation.

B. Evaluation Measures

BLEU [26], ROUGE [27], and METEOR [28] are key metrics for evaluating machine-generated text quality. BLEU uses n-gram precision, ROUGE measures overlapping text units, and METEOR considers synonyms and word order. These metrics are essential for evaluating translation, summarization, and question generation tasks.

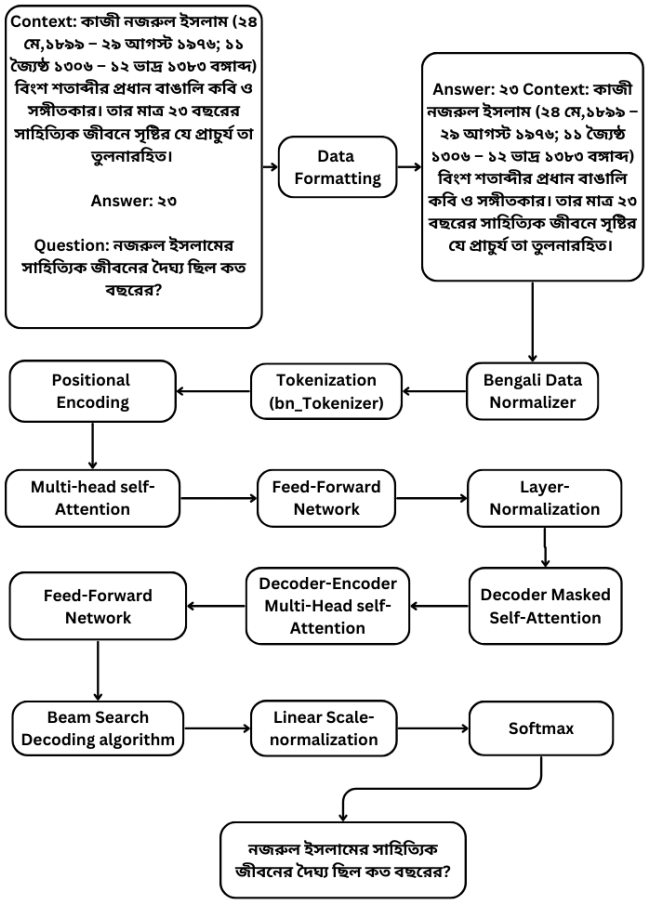

Fig. 4. Bengali Question Generation System: Data Processing and Model Architecture

C. Result

SAMPLE PREDICTION WITH BEAM SEARCH ALGORITHMS FOR PROPOSED METHOD

Context	Answer	Reference Question	Predicted Question
পূর্বে কলোরাডো মরুভূমি এবং কলোরাডো নদী অ্যারিজোনা সীমান্তে এবং নেভাদা রাজ্যের সীমান্তে মোজাভ মরুভূমি অবস্থিত। দক্ষিণে মেক্সিকো-যুক্তরাষ্ট্র সীমান্ত।	কলোরাডো মরুভূমি	অ্যারিজোনার সীমান্তে মরুভূমির নাম কি?	মোজাভ মরুভূমির পূর্বে কোন মরুভূমি অবস্থিত?
অ্যালগরিদম দ্বারা ব্যবহৃত সময় বা স্থানকে আবদ্ধ করে অনেক গুরুত্বপূর্ণ জটিল শ্রেণী নির্ধারণ করা যেতে পারে। এভাবে সংজ্ঞায়িত কিছু গুরুত্বপূর্ণ জটিল সিদ্ধান্ত সমস্যা নিম্নরূপ:	জটিল শ্রেণী	সময় এবং স্থান বা অনুরূপ পরিমাপের সীমা নির্ধারণ করতে প্রায়ই কি অ্যালগরিদম দ্বারা ব্যবহৃত হয়?	অ্যালগরিদম দ্বারা ব্যবহৃত সময় বা স্থানকে আবদ্ধ করে অনেক গুরুত্বপূর্ণ জটিল শ্রেণী কি নির্ধারণ করা যেতে পারে?
লস এঞ্জেলেস, অরেঞ্জ, সান দিয়েগো, সান বার্নার্ডিনো এবং রিভারসাইডের কাউন্টিগুলি রাজ্যের পাঁচটি সর্বাধিক জনবহুল এবং মার্কিন যুক্তরাষ্ট্রের শীর্ষ ১৫ টি জনবহুল কাউন্টিগুলির মধ্যে রয়েছে	লস এঞ্জেলেস	অরেঞ্জ, সান দিয়েগো, রিভারসাইড এবং সান বার্নার্ডিনো পাঁচটি কাউন্টির মধ্যে চারটি, সর্বশেষ কাউন্টিটির নাম কি?	অরেঞ্জ, সান দিয়েগো, সান বার্নার্ডিনো এবং রিভারসাইডের কাউন্টিগুলি কোথায় অবস্থিত?

Fig. 5. Sample Prediction with Beam Search Algorithm by the proposed method

In optimizing the Bangla Question Generation System, we compared Greedy and Beam decoding on BanglaT5 and MT5-base models. Greedy decoding ex-

celled for BanglaT5, while MT5-base with Beam decoding outperformed BanglaT5 across most metrics.

Our methodology adopts the MT5 model with a Bangla-specific tokenizer and text normalizer, paired with Beam Search decoding. This combination leverages MT5's language understanding and Beam Search's exploration of multiple question formulations, achieving high-quality, context-aware Bangla question generation.

TABLE II
Evaluation of question generation using finally selected models.

Models	Metric	b = 3	b = 5	b = 7
BanglaMT5	Rouge1	29.56	35.00	37.22
	Rouge2	12.54	16.41	16.98
	RougeL	17.29	31.00	30.29
	$\text{RougeL}_{\text{sum}}$	26.35	31.00	30.89
	SacreBLEU	11.45	14.14	14.20
	Meteor	0.2066	0.2192	0.2167
Proposed model	Rouge1	35.88	39.46	39.46
	Rouge2	16.20	17.68	17.83
	RougeL	36.01	37.02	36.72
	$\text{RougeL}_{\text{sum}}$	36.24	36.48	36.17
	SacreBLEU	9.48	9.56	9.57
	Meteor	0.2351	0.2358	0.2456

D. Performance Comparison

Our empirical performance analysis demonstrates the clear superiority of the proposed combination of bn-TOK model with mT5 base and beam search for Bengali question generation. This combination of models achieves the highest scores in key metrics, including Rouge (39.474), RougeL (36.725) and METEOR (0.245), indicating its exceptional performance in generating high-quality questions.

A previous study utilized BanglaT5 and mT5 transformer models with a greedy algorithm, omitting bnUnicodeNormalizer and bn-tokenizer, achieving a Rouge-1 score of 37.91 and BLEU-1 score of 38.57 with beam search (beam width 7). In contrast, our model surpasses these results with a Rouge-1 score of 39.474.

The Rouge score of 39.474 reflects the model's ability to capture a significant portion of the lexical overlap with the reference questions, suggesting that the generated questions are highly relevant and include many of the same words as the reference set. This indicates a high level of lexical accuracy and relevance. The RougeL score of 36.725 further underscores the model's effectiveness, as it suggests excellent preservation of the longest common sub-sequence. This implies that the generated questions maintain structural integrity and are syntactically well-aligned with the reference questions, highlighting the model's capacity to produce structurally coherent questions.

The METEOR score of 0.245, the highest among all models, highlights superior semantic adequacy and fluency. This demonstrates the model's ability to generate meaningful, coherent, and contextually appropriate questions while maintaining semantically rich and relevant sentence structures.

TABLE III
Comparison of Models with different features

Model	Metric	Score
Bert-TOK + IndicBart	Rouge	30.590
	RougeL	17.456
	secreBLEU	5.556
	METEOR	0.125
Bert-TOK + SahajBert	Rouge	31.225
	RougeL	24.256
	secreBLEU	8.264
	METEOR	0.157
Bert-TOK + Biswabangla	Rouge	32.041
	RougeL	24.922
	secreBLEU	9.439
	METEOR	0.147
Bert-TOK + BanglaBERT + Greedy	Rouge	30.134
	RougeL	25.275
	secreBLEU	8.052
	METEOR	0.186
mT5-TOK + mT5-base + Greedy	Rouge	36.850
	RougeL	28.613
	secreBLEU	8.125
	METEOR	0.192
Bn-TOK + BanglaT5 + Beam Search	Rouge	37.225
	RougeL	30.290
	secreBLEU	14.202
	METEOR	0.216
Bn-TOK + mT5-base + Beam Search	Rouge	39.474
	RougeL	36.725
	secreBBLEU	9.569
	METEOR	0.245

The SecreBLEU score of 9.569, though slightly lower than the bn-TOK + BanglaT5 + Beam Search combination, indicates strong syntactical and lexical accuracy. While improvements in fluency and lexical matching are possible, the proposed model excels overall, achieving the best balance of lexical overlap, structural accuracy, and semantic coherence.

Other model combinations, such as mT5-TOK + mT5-base + Greedy and bn-TOK + BanglaT5 + Beam Search, show strong but incomplete performances. The mT5-TOK + mT5-base + Greedy model achieves a Rouge score of 36.85 and a METEOR score of 0.192 but falls short in structural accuracy with a lower RougeL score of 28.613. Similarly, bn-TOK + BanglaT5 + Beam Search, with a Rouge score of 37.225 and SecreBLEU score of 14.202, excels in lexical and structural accuracy but lacks the semantic adequacy and fluency of the proposed model, as seen in its lower METEOR score of 0.216.

In conclusion, the bn-TOK + mT5-base + Beam Search model stands out as the most effective for generating high-quality Bangla questions. Its balanced performance across all critical metrics establishes it as the optimal choice for this task. The model's ability to generate questions that are lexically accurate, structurally coherent, and semantically meaningful underscores its superior performance, setting a benchmark for future research in Bangla Question Generation.

V. Limitations

But in the end, there were some limitations we faced. They are :

- The absence of a dedicated dataset for Bangla Question Generation (BQG) makes fine-tuning challenging, leading to potential accuracy issues and risks of hallucination.
- The unique architecture of the T5 model prevents the creation of hybrid models with other approaches.
- The consistency of generated questions when using synonyms remains untested.
- The trained model may reflect social biases, potentially producing toxic outputs.

VI. Conclusion

The study presents a Bengali question generation system using the encoder-decoder T5 model, which generates individual interrogative questions in an answer-aware style. The bn-TOK + mT5-base + Beam Search model stands out as the most effective for generating high-quality Bangla questions. Its balanced performance in all critical metrics establishes it as the optimal choice for this task. The model's ability to generate questions that are lexically accurate, structurally coherent, and semantically meaningful underscores its superior performance, setting a benchmark for future research in Bangla Question Generation.

References

[1] V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan, "The first question generation shared task evaluation challenge," 2010.

[2] M. Heilman and N. A. Smith, "Good question! statistical ranking for question generation," in Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics, pp. 609–617, 2010.

[3] T. Tahsin Mayeesha, A. Md Sarwar, and R. Rahman, "Deep learning based question answering system in Bengali," Journal of Information and Telecommunication, vol. 5, no. 2, pp. 145–178, 2021.

[4] A. Bhattacharjee, T. Hasan, K. Samin, M. Islam, M. Rahman, A. Iqbal, and R. Shahriyar, "Banglabert: Combating embedding barrier in multilingual models for low-resource language understanding," 2021. Edition: ArXiv Preprint ArXiv Pages: 2101 00204.

[5] M. Liu, V. Rus, and L. Liu, "Automatic chinese factual question generation," IEEE Transactions on Learning Technologies, vol. 10, no. 2, pp. 194–204, 2016.

[6] I. Kylliainen, "Neural Factoid Question Answering and Question Generation for Finnish," 2022.

[7] D. Kalpakchi and J. Boye, "Bert-based distractor generation for swedish reading comprehension questions using a small-scale dataset," arXiv preprint arXiv:2108.03973, 2021.

[8] F. Akyon, D. Cavusoglu, C. Cengiz, S. Altinuc, and A. Temizel, Automated question generation and question answering from Turkish texts using text-to-text transformers. ArXiv Preprint ArXiv: 2111.06476, 2021.

[9] V. Kumar, N. Joshi, A. Mukherjee, G. Ramakrishnan, and P. Jyothi, Cross-lingual training for automatic question generation. ArXiv Preprint ArXiv: 1906.02525, 2019.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, and D. Amodei, "Language Models Are Few-Shot Learners," 2020. Publication Title: In Proceedings of the 34th International Conference on Neural Information Processing Systems.

[11] A. Gatt and E. Krahmer, "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation," 2018.

[12] A. Ziegler, "GitHub Copilot research recitation," 2021.

[13] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung Survey of Hallucination in Natural Language Generation. ACM Comput. Surv, vol. 55, no. 12, 2023.

[14] R. Perera and P. Nand, "Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature," Computing and Informatics, vol. 36, pp. 1–31, 2017.

[15] T. Hasan, A. Bhattacharjee, M. Islam, K. Samin, Y.-F. Li, Y.-B. Kang, M. Rahman, and R. Shahriyar, XL-sum: Large-scale multilingual abstractive summarization for 44 languages. ArXiv Preprint ArXiv: 2106.13822, 2021.

[16] A. Kumar, H. Shrotriya, P. Sahu, R. Dabre, R. Puduppully, A. Kunchukuttan, A. Mishra, M. Khapra, and P. Kumar, "IndicNLG Benchmark: Multilingual Datasets for Diverse NLG Tasks in Indic Languages," 2022. Publication Title: In Conference on Empirical Methods in Natural Language Processing.

[17] A. Bhattacharjee, T. Hasan, W. Ahmad, K. Samin, M. Islam, A. Iqbal, M. Rahman, and R. Shahriyar, "BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla," 2022.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res, vol. 21, no. 140, pp. 1–67, 2020.

[19] A. Bhattacharjee, T. Hasan, W. Ahmad, and R. Shahriyar, BanglaNLG: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla. ArXiv Preprint ArXiv: 2205.11081, 2022.

[20] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," 2018.

[21] L. Dugan, E. Miltsakaki, S. Upadhyay, E. Ginsberg, H. Gonzalez, D. Choi, C. Yuan, and C. Callison-Burch, A feasibility study of answer-agnostic question generation for education. ArXiv Preprint ArXiv: 2203.08685, 2022.

[22] S. Raschka, "Understanding encoder and decoder." https://magazine.sebastianraschka.com/p/understanding-encoder-and-decoder, 2024. Accessed: 2024-09-29.

[23] R. Reddy, "Speech understanding systems: Summary of results of the five-year research effort at Carnegie-Mellon University," 1977.

[24] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," 2018. Edition: ArXiv Preprint ArXiv.

[25] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, The curious case of neural text degeneration. ArXiv Preprint ArXiv: 1904.09751, 2019.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," 2002. Pages: 311–318 Publication Title: In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

[27] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," 2004. Pages: 74–81 Publication Title: Text Summarization Branches Out,.

[28] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," 2005. Pages: 65–72 Publication Title: In Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.