BCoQA: Benchmark and Resources for Bangla Context-based Conversational Question Answering

Anonymous submission

Abstract

Developing a Bangla Context-based Conversational Question Answering (CCQA) system presents unique challenges, including limited domain-specific data, inadequate translation methods, and a lack of pretrained language models. In this work, we address these obstacles by constructing a robust Bangla CCQA dataset through quality controlled machine translation and LLM based augmentation of established English CCQA datasets, followed by partitioning into training, validation and test splits. We finetune and then evaluate the performance of various existing sequence-to-sequence models using the train and test split respectively, by appending conversation history into the input prompt to preserve context. The entire dataset and the testing script have been made publicly available on GitHub for benchmarking future models. This initiative marks a significant step in advancing conversational AI for Bangla, setting a foundation for further research and development in the field. All the code and resources are available through this github link.

1 Introduction

Context-based Conversational Question Answering (CCQA) systems facilitate a natural flow of dialogue by understanding and generating responses that align with the context of a conversation. While significant strides have been made in developing CCQA systems for resource-rich languages, Bangla, a language spoken by over 300 million people, has not yet seen advancements in this area. The absence of a CCQA system that can handle contextual question-answering and maintain a conversational flow in Bangla poses a critical gap in the field of natural language processing (NLP). To address this gap, we introduce BCoQA, a novel dataset and benchmarking suite that emulates the success of CoQA (Reddy et al., 2018), a pioneering conversational question answering challenge for English. Our work aims to bridge the gap in Bangla CCQA systems by providing a high-quality dataset that captures the nu-

ইন্টেল কর্পোরেশন (এছাড়াও ইন্টেল নামে পরিচিত, ইন্টেল হিসাবে শৈলীকৃত) একটি আমেরিকান বহুজাতিক কর্পোরেশন এবং প্রযুক্তি কোম্পানি যার সদর দপ্তর ক্যালিফোর্নিয়ার সান্তা ক্লারায় অবস্থিত। এটি স্যামসাং দ্বারা অধিগৃহীত হয়ে রাজস্বের উপর ভিত্তি করে বিশ্বের দ্বিতীয় বৃহত্তম এবং দ্বিতীয় সর্বোচ্চ মূল্যবান অর্ধপরিবাহী চিপ নির্মাতা এবং এক্স৮৬ সিরিজের মাইক্রোপ্রসেসরের উদ্ভাবক, যা বেশিরভাগ ব্যক্তিগত কম্পিউটারে (পিসি) পাওয়া যায়। ইন্টেল কম্পিউটার সিস্টেম প্রস্তুতকারক যেমন অ্যাপল, লেনোভো, এইচপি, এবং ডেল এর জন্য প্রসেসর সরবরাহ করে। ইন্টেল মাদারবোর্ড চিপসেট, নেটওয়ার্ক ইন্টারফেস কন্ট্রোলার এবং ইন্টিগ্রেটেড সার্কিট, ফ্ল্যাশ মেমোরি, গ্রাফিক্স চিপ, এমবেডেড প্রসেসর এবং যোগাযোগ ও কম্পিউটিং সম্পর্কিত অন্যান্য ডিভাইস তৈরি করে। ইন্টেল কর্পোরেশন ১৯৬৮ সালের ১৮ জুলাই অর্ধপরিবাহী অগ্রগামী রবার্ট নয়েস এবং গর্ডন মুর দ্বারা প্রতিষ্ঠিত হয় ...

Q_1: এই প্রবন্ধের বিষয়বস্তু কী?
A_1: ইন্টেল কর্পোরেশন

Q_2: কোম্পানির সদর দপ্তর কোথায়?
A_2: সান্তা ক্লারা, ক্যালিফোর্নিয়া

Q_3: তারা কি একটি বহুজাতিক কর্পোরেশন?
A_3: হ্যাঁ

Q_4: ইন্টেল কি আবিষ্কার করেছে?
A_4: এক্স৮৬ সিরিজ মাইক্রোপ্রসেসর

Q_5: এটি কিসে ব্যবহৃত হয়?
A_5: অধিকাংশ ব্যক্তিগত কম্পিউটারে (পিসি)

Q_6: কোম্পানি টি কখন প্রতিষ্ঠিত হয়েছিল?
A_6: জুলাই ১৮, ১৯৬৮

Q_7: একজন প্রতিষ্ঠাতার নাম বলুন।
A_7: রবার্ট নয়েস

Q_8: আর কেউ?
A_8: গর্ডন মুর

Q_9: তিনি আর কি প্রতিষ্ঠা করেন?
A_9: অজানা

Figure 1: A conversation from the BCoQA dataset showing entity of focus in colors.

Dataset	Conversational	Answer Type
SQuAD_bn (Bhattacharjee et al., 2021)	✗	Spans, Unanswerable
BanglaRQA (Ekram et al., 2022)	✗	Spans, Yes/No, Unanswerable
Tydiqa (Clark et al., 2020)	✗	Spans, Yes/No
QAmeleon (Agrawal et al., 2022)	✗	Free-form Text
BCoQA (this work)	✓	Free-form text, Unanswerable

Table 1: Comparison of BCoQA with existing Bangla reading comprehension datasets.

ances of human conversations in Bangla. The BCoQA dataset is crafted with 3 objectives:

1. Mimic the natural flow of questions in Bangla conversations, where each question builds upon the previous one (Figure 1). For example, Q7 (একজন প্রতিষ্ঠাতার নাম বলুন।) requires conversation history to answer.

2. Ensure natural answers in conversations, using free-form answers instead of limited text spans. For instance, Q3 (তারা কি একটি বহুজাতিক কর্পোরেশন?) has no span-based answer.

3. Include unanswerable questions, which are very common in conversational question answering. For example, Q9 (তিনি আর কি প্রতিষ্ঠা করেন?) has no answer in the passage.

By introducing BCoQA, we provide the necessary stepping stone for Bangla Question Answering systems to evolve to a more natural state. Our dataset and benchmarking suite offer a foundation for the community to build upon, enabling the development of more advanced and robust conversational systems for the Bangla language. We also benchmark several state-of-the-art sequence-to-sequence models to provide a baseline for future research in this area, achieving an F1 score of 46.9%. In contrast, humans achieve 78.5% F1, indicating that there is significant room for improvement.

1.1 Task Definition

The goal is to answer the current question in conversation, considering the passage and conversation history. If the answer can't be found, the output should be "অজানা". Figure 1 shows how the entity of focus[1] changes throughout the conversation.

[1]a series of pronouns or noun phrases that refer to the same entity or concept in a conversation or text

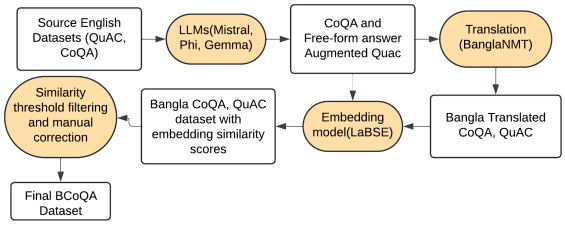

Figure 2: BCoQA Dataset Creation Workflow

Domain	Conversations
Children's Stories	429
Literature	662
Mid/High School Stories	1384
News	1264
Wikipedia Articles	10430
Total	14169

Table 2: Domain distribution in the final BCoQA dataset, showing the over-representation of Wikipedia articles due to QuAC's singular passage source.

2 Dataset Creation

2.1 Source Dataset Selection

Creating a full Bangla dataset from scratch is not feasible due to limited resources. Instead, we translated existing English context-based conversational question-answering datasets. We selected two datasets: CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018). Both share common features: context passages, multi-turn conversations, unanswerable questions, and answer spans as evidence. However, they differ in collection methods, leading to some key differences. QuAC provides only answer spans, whereas CoQA has a diverse range of domains. QuAC questions are more open-ended, and its domains are limited to Wikipedia articles, resulting in an unbalanced domain distribution in the combined dataset

Context: In 1969, still in the Pre-Crisis continuity, writer Dennis O'Neil and artist Neal Adams return Batman to his darker roots. One part of this effort is writing Robin out of the series by sending Dick Grayson to Hudson University and into a separate strip in the back of Detective Comics. The by-now Teen Wonder appears only sporadically in Batman stories of the 1970s as well as a short lived revival of The Teen Titans. In 1980, Grayson once again takes up the role of leader of the Teen Titans, now featured in the monthly series The New Teen Titans, which became one of DC Comics's most beloved series of the era. During his leadership of the Titans, however, he had a falling out with Batman, leading to an estrangement that would last for many years.

Question: What role did he play in Teen Titans?

Answer Span: In 1980, Grayson once again takes up the role of leader of the Teen Titans,

Generated free-form answer: He played the role of leader in Teen Titans.

Figure 3: Example of converting answer spans into free-form answers using LLMs.

(Table 2).

2.2 Creating Free-Form Answers

CoQA already includes free-form answers along with the answer spans. Since QuAC only provides answer spans, we used large language models (LLMs) to generate free-form answers. We input the context, question, and answer evidence span into three open-source LLMs: Mistral 7B (Jiang et al., 2023) (Accessed 25th December, 2023), Phi-2 2.7B by Microsoft(Accessed 30th December, 2023), and Gemma 7B by Google(Accessed 25th February, 2024). After evaluating their performance, we chose Mistral 7B for its consistent and reliable answers. Figure 3 shows an example of a Mistral 7B-generated free-form answer.

2.3 Machine Translation & Quality Assurance

To ensure high-quality translations, we followed a rigorous curation process similar to the one used in the squad_bn dataset (Bhattacharjee et al., 2021). We translated CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018) into Bangla using a state-of-the-art English-to-Bangla translation model (Hasan et al., 2020). However, Despite being the state-of-the-art, Bangla-to-English machine translation models still struggle with accuracy, with even the best-performing model achieving a SacreBLEU score of only 22.0 (Hasan et al., 2020). To minimize errors, we employed Language-Agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2020) to measure the semantic similarity between translated Bangla sentences and their original English counterparts. We computed the cosine distance between the sentence embeddings, which served as a similarity score. This approach allowed us to quantitatively evaluate the translation quality. We settled on a similarity score threshold of 0.7[2] for the training set. For the validation and test sets, we used a stricter threshold of 0.88[3] to ensure the model was evaluated on the most accurate translations. Prior to translation, we conducted extensive data cleaning on the original English datasets to remove noise, inconsistencies, and potential sources of translation errors. After completing the entire process, we removed approximately 24% of the conversations and split the remaining data into training, validation, and testing sets, ensuring a similar mix of answer types in each group. The final dataset structure is shown in Table 3.

3 Benchmarking Existing Models

3.1 Models

We treated the CCQA task as a conversational response generation problem. We combined the conversation history with the prompt in a consistent way during training and testing. We didn't use reading comprehension models like BanglaBERT (Bhattacharjee et al., 2022) as they aren't suitable for generating free-form answers. We fine-tuned existing sequence-to-sequence (seq2seq) models that support Bangla on our training data and evaluated their performance on the test set. We prepared the input by concatenating the passage, conversation history, and current question in a specific format: P <q> Q_1 <a> A_1 ... <q> Q_{i-1} <a> A_{i-1} <q> Q_i <a>, where P is the context passage, <q> and <a> are the question and answer markers respectively.

[2]Analyzed the frequency distribution of similarity scores and manually inspected translation quality at various score levels.

[3]Increased the similarity score until we had just enough conversations to create adequately sized test and validation sets.

Split Name	Data points/Conversations	Yes/No	Unknown	Short	Long
Train	13169	32332	14087	34816	57505
Validation	500	1138	380	1385	1833
Test	500	1128	389	1397	1885
Total	14169				

Table 3: Dataset split analysis with different answer types

Model	Parameter Count	Exact Match	F1 Score
Human Performance	—	72.1	78.5
banglat5 (Bhattacharjee et al., 2022)	248M	35.3	46.9
mt5-base (Xue et al., 2020)	582M	32.2	41.7
banglat5_small (Bhattacharjee et al., 2022)	60.5M	29.9	40.2
mbart-large-50 (Lewis et al., 2019)	611M	31.9	39.9

Table 4: Average performance scores for Human and Models on test data

Model	Yes/No EM	Unknown EM	Short F1	Long F1
bt5	75.9	28.8	**50.3**	**30.8**
mt5	**77.7**	13.1	43.0	24.9
bt5_sm	71.6	15.4	39.9	26.6
mbart	75.8	**33.4**	39.5	20.1

Table 5: Model scores for different question-answer types

3.2 Evaluation

We used the macro average F1 score of word overlap as our primary evaluation metric, consistent with CoQA. We assessed our model's performance by predicting the next answer using the gold standard answers from the conversation history. To ensure a fair evaluation, we removed punctuation and stop words (e.g., pronouns, verbs, conjunctions) from both the gold and predicted answers. We established a baseline by evaluating human performance on the test set. We recruited 10 participants to respond to 50 conversations each, providing them with the context, questions, and conversation history. We compared the results to the models' performance. Table 4 shows the test data results, with the top-performing model, banglat5, achieving an Exact Match score of 35.3 and an F1 Score of 46.9. However, it still falls short of human performance by 31.6 points in F1 score. Notably, banglat5 outperforms other models despite having fewer parameters (223M) due to its pretraining on the Bangla corpus. Similarly, the smaller version of banglat5, with about 1/10th the number of parameters(60.5M), holds its own against larger multilingual models like mt5-base and mbart-large-50, which have 582M and 611M parameters, respectively. Upon closer examination of the results shown in table 5, we observe that the models perform best on yes/no type questions, which is a expected phenomenon for seq2seq models(see (Feng et al., 2020)). This is because YES/NO answers often rely on simple factual information or binary decisions, making it easier for the models to predict the correct response. The banglat5 variants also excel in providing accurate long answers, indicating that Bangla pretraining is essential for generating longer responses.

4 Conclusion

We introduce BCoQA, a large-scale dataset for developing context-based conversational question answering systems in Bangla. By leveraging state-of-the-art language models and machine translation systems, we aim to bridge the gap between Bangla and more established languages like English. Our experiments highlight the challenges faced by resource-scarce languages like Bangla, where even the best-performing models trail human performance by a substantial margin. However, these models show competitive performance against larger multilingual models, demonstrating the potential of tailored approaches for Bangla. We hope this work will inspire further research in conversational modeling and NLP advancements in Bangla, enabling more natural and effective human-machine communication.

Limitations

Our approach to creating and evaluating BCoQA has several limitations. Firstly, our use of machine translation for the entire dataset may have introduced errors and unnatural phrasing, despite our quality thresholding efforts. While the translations are semantically meaningful, they may not be entirely natural-sounding, which posed a learning curve for human performance evaluators. However, despite this limitation, the models were able to generalize well enough to Bangla texts that they performed competently when tested with real-world, non-translated Bangla contexts and natural conversations. This suggests that while the translated dataset may not be perfect, it is still a valuable resource for training CCQA systems that can generalize to real-world scenarios.

Secondly, the domain distribution of our dataset is heavily biased towards Wikipedia articles, which may not be representative of all possible conversational question answering scenarios. To mitigate this, we maintained equal domain distribution in the test set to ensure a fair assessment of CCQA systems. However, this bias in training set may still impact the generalizability of our dataset.

Thirdly, we only evaluated sequence-to-sequence models, which may not be the most effective architectures for CCQA tasks. In fact, the CoQA paper (Reddy et al., 2018) suggests that seq2seq models perform poorly compared to other architectures. Our decision to only provide free-form answers and not answer spans may have limited the types of models that can be evaluated on our dataset. While we believe that text-to-text modeling is the future of language modeling, it is unclear how other model architectures may perform on our dataset. Future work is needed to explore the performance of other models on BCoQA.

References

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2022. Qameleon: Multilingual qa with only 5 examples.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2022. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac : Question answering in context.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. Banglarqa: A benchmark dataset for under-resourced bangla language reading comprehension-based question answering with diverse question-answer types. pages 2518–2532. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Google. 2024. Gemma 7b.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. pages 2612–2623. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Microsoft. 2023. Phi-2 2.7b.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer.

A English Examples

Here are the Bangla context-conversation examples, translated in English.

Intel Corporation (also known as Intel, stylized as intel) is an American multinational corporation and technology company headquartered in Santa Clara, California. It is the world's second largest and second highest valued semiconductor chip makers based on revenue after being overtaken by Samsung, and is the inventor of the x86 series of microprocessors, the processors found in most personal computers (PCs). Intel supplies processors for computer system manufacturers such as Apple, Lenovo, HP, and Dell. Intel also manufactures motherboard chipsets, network interface controllers and integrated circuits, flash memory, graphics chips, embedded processors and other devices related to communications and computing. Intel Corporation was founded on July 18, 1968, by semiconductor pioneers Robert Noyce and Gordon Moore. ...

Q_1: What is the subject of the article?
A_1: Intel Corporation

Q_2: Where is the company's headquarters?
A_2: Santa Clara, California

Q_3: Are they a multinational company?
A_3: Yes.

Q_4: What did Intel invent?
A_4: x86 series of microprocessors

Q_5: Where is it used?
A_5: Most personal computers (PCs)

Q_6: When was the company founded?
A_6: July 18, 1968

Q_7: Name One Founder.
A_7: Robert Noyce

Q_8: And the other?
A_8: Gordon Moore

Q_9: What else did he establish?
A_9: Unknown

Figure 4: A conversation from the BCoQA dataset showing coreference chains in colors - Translated from figure 1

B Finetuning Setup

Finetuning Setup We finetuned our models on the BCoQA dataset using the Seq2SeqTrainer from the Huggingface transformers library. The finetuning setup consisted of:

- 2 epochs of training
- Learning rate of 4e-5
- Maximum sequence length of 1024
- Adafactor optimizer for BanglaT5, BanglaT5 Small, and MT5. AdamW optimizer for MBART.
- Batch size between 4-8 depending on the model size to maximize throughput.

Our finetuning experiments were run on a PC setup equipped with an NVIDIA RTX 4090 GPU.