https://doi.org.10.3126/batuk.v12i1.90049
THE BATUK : A Peer Reviewed Journal of Interdisciplinary Studies
Vol. 12, Issue No. 1, Jan 2026, Page: 111-120
ISSN 2392-4802 eISSN 2565-4934
Received: 9 Dec 2025 Revised: 21 Dec 2025 Accepted: 22 Jan 2026

Comparative Analysis of Transformer Model: mBART and mT5 on Question Answering System for Nepali Text

Raju Shrestha[1], Krisha Shrestha[2], Basant Karki[3]
Asian College of Higher Studies

Abstract

Despite significant advances in English question answering using transformer models such as Text-To-Text Transfer Transformer (T5), Bidirectional Auto-Regressive Transformers (BART), and Generative Pre-trained Transformer (GPT) trained on datasets like Standford Question Answering (SQuAD), research on Nepali question answering remains limited due to the scarcity of annotated data and fine-tuned models. This study presents a comparative analysis of two multilingual transformer models mBART and mT5 for Nepali question answering using transfer learning. A translated Nepali SQuAD dataset was developed and fine-tuned with both models, incorporating data augmentation to address data scarcity. Evaluation using BLEU, ROUGE, BERTScore, Exact Match, and F1 Score shows that both models perform well, with mBART slightly outperforming mT5. This work provides a foundation for future research on Nepali question answering systems.

Keywords: question answering, standard question answer dataset, multilingual transformers, mBART, mT5

Introduction

Considering the growing Nepal's growing digital infrastructure, Nepali-language Question Answering (QA) systems have strong potential for applications in government, legal, healthcare, education, and agriculture sectors. While transformer-based QA systems such as BART, T5, and GPT have achieved significant success in English using large datasets like SQuAD, progress in Nepali QA remains limited due to the scarcity of large-scale, well-annotated datasets and fine-tuned transformer models. This lack of linguistic resources continues to hinder the development of effective Nepali QA systems.

[1] Correspondence concerning this article should be addressed to Raju Shrestha, Lecturer, Department of CSIT, Asian College of Higher Studies. Email: raju@achsnepal.edu.np
[2] Lecturer, Department of CSIT. Email: krisha@achsnepal.eu.np
[3] Lecturer, Department of CSIT. Email: basantajung@achsnepal.edu.np

This research focuses on Nepali question answering using multilingual transformers mBART and mT5, fine-tuned on a custom Nepali SQuAD-style dataset containing passages and multiple question-answer pairs. QA systems can be closed-domain, open-domain, extractive, or generative (Devlin et al., 2019; Ferrucci, 2012; Lewis et al., 2020; Manning et al., 2020), and models like mBART, mT5, and GPT have significantly improved performance, particularly in multilingual and low-resource language settings (Budur et al., 2024).

The primary objective of this research paper is to evaluate and compare the performance of multilingual transformer models, mBART and mT5, for Nepali question answering. Specifically, to implement these models for Nepali QA and assess their performance using standard evaluation metrics, including BLEU, ROUGE, BERTScore, Exact Match, and F1 Score.

Literature Review

QA has advanced from rule-based and information retrieval models to end-to-end neural network architectures over the past two decades. Transformer-based deep learning models have made QA systems more accurate, context-aware, and linguistically robust. Early rule-based systems relied on linguistic tools like POS tagging, NER, and syntactic parsing, performing well in constrained domains but struggling with scalability, domain adaptation, and low-resource languages like Nepali (Abacha &Zweigenbaum, 2015).

The advent of statistical methods and word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe(Pennington et al., 2014) enabled deep learning models to capture textual semantics, improving performance on NLP tasks, including QA (Mikolov et al., 2013). Models such as BiDAF(Seo et al., 2016) and DrQA(Chen et al., 2017) benefited from these embeddings, achieving strong results on datasets like SQuAD(Rajpurkar et al., 2016), but required large task-specific data—challenging for low-resource languages. Transfer learning with ULMFiT(Howard & Ruder, 2018), ELMo(Peters et al., 2018), and BERT (Devlin et al., 2019) allowed pre-trained models to be fine-tuned on smaller QA datasets, enhancing performance even in low-resource languages like Marathi (Amin et al., 2023).

The Transformer architecture (Vaswani et al., 2017) revolutionized NLP by replacing RNNs and CNNs with self-attention, efficiently capturing long-range dependencies. This foundation led to state-of-the-art models like BERT, RoBERTa, and GPT, supporting transfer learning where large-scale pretraining is fine-tuned on smaller, task-specific datasets for improved contextual understanding and efficiency (Vaswani et al., 2017).

In QA, extractive models like BERT achieved state-of-the-art results on SQuAD(Rajpurkar et al., 2016) but were limited for generative tasks. Encoder-decoder models such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) became popular for both extractive and generative QA, producing fluent, contextually relevant answers rather than merely extracting text spans.

For low-resource languages like Nepali, multilingual models such as mBERT, XLM-R, and mBART(Liu et al., 2020) enable zero- or few-shot transfer across languages. mBART, supporting over 50 languages, excels in both extractive and generative QA via cross-lingual transfer (Lewis et al., 2020). Despite limited Nepali QA datasets, translated resources like SQuAD and pretrained multilingual models facilitate effective Nepali QA systems (Tang et al., 2020).

mBART (Multilingual BART) is a sequence-to-sequence denoising autoencoder supporting over 50 languages, including Nepali (mBART50). Unlike encoder-only models like mBERT, its encoder-decoder architecture is well-suited for generative QA, combining understanding and answer generation. Fine-tuning mBART on low-resource QA datasets leverages knowledge from high-resource languages, yielding promising results (Liu et al., 2020).

mT5, the multilingual version of T5, is a text-to-text encoder-decoder model trained on mC4 (multilingual Colossal Clean Crawled Corpus), a large-scale corpus covering 101 languages (Xue et al., 2021). Its Transformer-based architecture captures multilingual semantics and structure, enabling it to generate responses for generative QA tasks in low-resource languages (Xue et al., 2021).

Although spoken by over 40 million people, Nepali differs syntactically from English and remains underrepresented in NLP research (Gautam et al., 2022). Most Nepali NLP work focuses on POS tagging, sentiment analysis, and machine translation, with few studies on QA due to limited annotated datasets and pre-trained models.

Recent efforts, including the FLORES dataset and multilingual benchmarks from Facebook AI and Hugging Face, now include Nepali, enabling better evaluation and fine-tuning of models like mBART(Goyal et al., 2022).

Multilingual models fine-tuned on translated or synthetic QA datasets can achieve strong performance, especially with semantic-aware metrics like BERTScore(Zhang et al., 2020). Transformer models such as mBART offer a promising approach for QA in low-resource languages like Nepali (Liu et al., 2020).

Methodology

Data Collection

This study used the Nepali version of the Stanford Question Answering Dataset (SQuAD) obtained from Hugging Face. The dataset is a translated form of the original English SQuAD and contains 19,048 Nepali contexts with multiple question–answer pairs, formatted in JSON for compatibility with transformer-based models such as mBART and mT5. The dataset is used to evaluate and compare the performance of these models on Nepali question answering.

Context**:** "मुख्यभवनकोसुनकोगुम्बदकोमाथिरानीमरियमकोसुनकोप्रतिमाछ।मुख्यभवनकोछेउमारयसकोछेउमा, ""विनिटएडमेओमनेस"" सँगarms upraised with the legend ""Venite Ad Me Omnes"" कोप्रतिमाख्रीष्टकोछ।मुख्यभवनकोछेउमापवित्रहृदयकोbasilica छ।basilicaकोछेउमाग्रोटोछ, म्यारिअनप्रार्थनारप्रतिबिम्बकोस्थानायोLourdes, फ्रान्समाग्रोटोकोप्रतिमाहोजहाँरानीमरियमलेसन्१८५८मासेन्टबर्नाडेटसुबिरोसलाईदेखापथ्र्यो।मुख्यड्राइभकोअन्त्यमा (रतीनप्रतिमाहरूरसुनकोगुम्बदहुँदैजडानगर्नेप्रत्यक्षरेखामा) मरियमकोसरल, आधुनिकढुङ्गाकोप्रतिमाछ।"

Table 1

Question Answer Dataset

S.N.	Question	Answer
1.	सन् १८५८ मा लोर्डर फ्रान्समा जम्मैले भर्जिन मरियम देखाए भन्ने आरोप लगाइएको थियो?	सेन्ट बर्नाडेटे सुबिरोस
2.	Notre Dame मुख्य भवनको अगाडि के छ?	ख्रीष्टकोप्रतिमा
3.	Notre Dameको पवित्र हृदयको basilica को संरचना छेउमा छ?	मुख्य भवन
4.	Notre Dameको गुफा के हो?	मारियन प्रार्थना र विचारको स्थान
5.	Notre Dameको मुख्य भवनको माथिल्लो भागमा के बस्छ?	Virgin Mary को सुनौलो प्रतिमा

Data Processing

Tokenization: Performed using the tokenizer associated with mBART (mbart50Tokenizer) and mT5 (mT5Tokenizer), ensuring consistency with the pre-trained vocabulary.
Data Cleaning: Removed noise, normalized Nepali text, and verified alignment between questions and answers.
Formatting: Prepared the dataset in a format compatible with mBART and mT5 training (context + question → answer).
*Data Splitting***:** The cleaned dataset is divided into training and validation sets, typically in the ratio 80% and 20% to evaluate model performance.

Model Selection and Fine-Tuning

Multi-lingual Bidirectional and Auto Regressive Transformer (mBART).mBART is a multilingual extension of BART pretrained on large-scale corpora covering 50 languages, including Nepali, using a Transformer-based encoder–decoder architecture (Liu et al., 2020). It employs a denoising autoencoder objective with noise injection and sequence reconstruction, enabling strong cross-lingual transfer and making it well suited for sequence-to-sequence tasks such as question answering (Liu et al., 2020).

Multi-lingual Text-to-Text Transfer Transformer (mT5).mT5 extends the T5 framework to 101+ languages, including Nepali, and is pretrained on the multilingual Colossal Clean Crawled Corpus (mC4), enabling strong cross-lingual generalization and effective transfer learning for low-resource languages (Gautam et al., 2022).

Model Setup. Both models were fine-tuned using a parallel setup for consistency and fair evaluation.The model architecture for mBART and mT5 is given in Table 2.

Table 2

Model Architecture for mBART and mT5

Features	mBART50	mT5 (Small)
Architecture	Encoder-Decoder	Encoder-Decoder
Pretraining Task	Denoising Autoencoding	Text-to-Text(Span-Corruption)
Parameters	~610M (mbart-large-50)	~300M
Tokenizer	SentencePiece	SentencePiece
Pretrained on	50 languages	101 languages
Language ID Token	Required (eg.<ne_NP>)	Not required

Tokenization. The mBART50 uses a language-specific token (<ne_NP>) for Nepali, while mT5 does not require language tokens. Accordingly, Fine Tuning Parameters are presented in Table 3.

Table 3

Fine-Tuning Configuration

Parameter	mBART	mT5
Max input length	256	256
Maxoutput length	64	64
Batch size	4	4
Epochs	3	5
Learning rate	5e-5	3e-5
Optimizer	AdamW	AdamW
Scheduler	linear	linear

Hyper-parameter Configuration. Key hyperparameters were tuned for stability and efficiency. mBART-large-50 uses 12 encoder/decoder layers (~610M parameters), while mT5-small has 8 layers (~300M parameters); both employ shared embeddings and multi-head attention. Training used masked loss computation (padding tokens = −100) to ensure accurate gradient updates.

Evaluation Metrics. Given the generative nature of the QA system, multiple metrics are used to assess both syntactic and semantic quality: BLEU and ROUGE-L for n-gram overlap, BERTScore for semantic similarity, and Exact Match (EM) with F1-score for token-level accuracy and overlap.

Results

Based on the results drawn, a fair comparison of mBART50 and mT5-small for Nepali question answering, with both models trained and evaluated under identical experimental settings is made.

Table 4

Training and Validation Loss

Metric	mBART50	mT5-small
Best Training Loss	0.0765	0.2442
Best Validation Loss	0.2080	0.2574
Training Accuracy	98.22%	95.27%
Validation Accuracy	96.55%	95.36%

As shown in Table 4, mBART50 outperformed mT5-small with lower training (0.0765 vs. 0.2442) and validation loss (0.2080 vs. 0.2574), indicating faster convergence and better generalization. It also achieved higher training (98.22% vs. 95.27%) and validation accuracy (96.55% vs. 95.36%), reflecting a stronger ability to learn and generalize Nepali text.

Accordingly, evaluation using BLEU, ROUGE-L, BERTScore, Exact Match, and F1 Score shows mBART50 consistently outperforming mT5-small. BLEU (0.1738 vs. 0.1370) and ROUGE-L (0.0345 vs. 0.0202) indicate better n-gram overlap and coverage of relevant context. BERTScore (0.8932 vs. 0.8802) reflects superior semantic alignment, while Exact Match (24.73% vs. 19.10%) and F1 Score (39.30 vs. 32.75) demonstrate higher precision, recall, and overall answer quality (Table 5). These results highlight mBART50's stronger performance in Nepali QA.

Table 5

Evaluation Metrics

Metric	mBART50	mT5-small
BLEU Score	0.1738	0.1370
ROUGE-L	0.0345	0.0202
BERTScore	0.8932	0.8802
Exact Match	24.73	19.10
F1 Score	39.30	32.75

Discussion

Results of the study clearly shows that mBART50 significantly outperformed mT5-small across all evaluation metrics. The several factors contributed to this superior performance are:

mBART50's superior performance is attributed to its multilingual pretraining, which provides robust linguistic understanding, its denoising autoencoder architecture suited for sequence-to-sequence QA tasks, and its larger model capacity, enabling it to capture complex patterns in Nepali text. In contrast, mT5-small's smaller size and simpler training limited its ability to match mBART50 in accuracy, loss reduction, and evaluation metrics.

The sample test generated by two different transformer models; mBART and mT5 on question answering system for Nepali text are shown in Figure 1 and 2.

Figure 1

Test Sample Generated by mBART

```
# Inference
context = "नेपालको संविधान २०७२ सालमा जारी गरिएको थियो। यसमा संघीयता, धर्म निरपेक्षता लगायतका प्रावधान छन्।"
question = "नेपालको संविधानमा के प्रावधान छन्?"
input_text = f"प्रश्न: {question} सन्दर्भ: {context}"

inputs = tokenizer(input_text, return_tensors="pt", truncation=True, padding=True).to(device)
output_ids = model.generate(**inputs, max_length=64)
output = tokenizer.decode(output_ids[0], skip_special_tokens=True)

print("🔍 उत्तर:", output)
```

```
🔍 उत्तर: संघीयता, धर्म निरपेक्षता
```

Figure 2

Test Sample Generated by mT5

```
# Inference
context = "नेपालको संविधान २०७२ सालमा जारी गरिएको थियो। यसमा संघीयता, धर्म निरपेक्षता लगायतका प्रावधान छन्।"
question = "नेपालको संविधानमा के प्रावधान छन्?"
input_text = f"प्रश्न: {question} सन्दर्भ: {context}"

inputs = tokenizer(input_text, return_tensors="pt", truncation=True, padding=True).to(device)
output_ids = model.generate(**inputs, max_length=64)
output = tokenizer.decode(output_ids[0], skip_special_tokens=True)

print("🔍 उत्तर:", output)
```

```
🔍 उत्तर: संघीयता, धर्म निरपेक्षता
```

Conclusion

Experimental results show that mBART50 outperformed mT5-small across nearly all metrics. mBART50 achieved higher training and validation accuracy (98.22% and 96.55% vs. 95.27% and 95.36%) and lower training and validation loss (0.0765 and 0.2080 vs. 0.2442 and 0.2574), indicating better convergence and generalization. Standard NLP metrics also favored mBART50, with higher BLEU (0.1738 vs. 0.1370), ROUGE-L (0.0345 vs. 0.0202), BERTScore (0.8932 vs. 0.8802), Exact Match (24.73 vs. 19.10), and F1 Score (39.30 vs. 32.75), demonstrating superior answer quality. These results suggest mBART50 is better suited for Nepali QA, likely due to its multilingual encoder-decoder architecture and pretraining aligned with Nepali language structure.In conclusion, both models show the potential of multilingual transformers for low-resource languages, but mBART50 is more effective for building high-quality Nepali QA systems.

References

Abacha, A. B., &Zweigenbaum, P. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. *Information Processing & Management, 51*(5), 570–594. https://doi.org/10.1016/j.ipm.2015.04.006

Amin, D., Govilkar, S., & Kulkarni, S. (2023). Question answering using deep learning in low resource Indian language Marathi. *arXiv*. https://arxiv.org/abs/2309.15779

Budur, E., Ozcelik, R., Soylu, D., Khattab, O., Gungor, T., & Potts, C. (2024). *Building efficient and effective OpenQA systems for low-resource languages*. arXiv. https://arxiv.org/abs/2401.03590

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1870–1879.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://arxiv.org/abs/1810.04805

Ferrucci, D. A. (2012). Introduction to *This is Watson*. *IBM Journal of Research and Development, 56*(3), 1–15. https://doi.org/10.1147/JRD.2012.2184356

Gautam, M., Timilsina, S., & Bhattarai, B. (2022). NepBERT: Nepali language model trained on a large corpus. *Proceedings of the Association for Computational Linguistics*, 273–284.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M. A., Guzman, F., & Fan, A. (2022). The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 522–538.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv*. https://arxiv.org/abs/1801.06146

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://arxiv.org/abs/1910.13461

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics, 8(1)*, 726–742.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://arxiv.org/abs/1907.11692

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, 117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://arxiv.org/abs/1301.3781

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of NAACL-HLT*, 2227–2237.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *arXiv*. https://arxiv.org/abs/1606.05250

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. *arXiv*. https://arxiv.org/abs/1611.01603

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., & Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv*. https://arxiv.org/abs/2008.00401

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., &Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv*. https://arxiv.org/abs/2010.11934

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations*.