Received Date:5th February, 2025
Revision Date: 12th February, 2025
Accepted Date: 6th March, 2025

Extractive Nepali Question Answering System

Yunika Bajracharya[1]*, Suban Shrestha[2], Saurav Bastola[3], Sanjivan Satyal[4]

[1]Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, TU, Nepal. Email: 076bct095.yunika@pcampus.edu.np
[2]Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, TU, Nepal. Email: 076bct082.suban@pcampus.edu.np
[3]Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, TU, Nepal. Email: 076bct073.saurav@pcampus.edu.np
[4]Assoc. Professor, Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, TU, Nepal.
Email: sanziwan.satyal@pcampus.edu.np

Abstract — There is a noticeable gap in language processing tools and resources for Nepali, a language spoken by more than 17 million people [1] yet significantly underrepresented in computational linguistics. We present an Extractive Nepali Question Answering System designed to generate precise, contextually accurate responses in Nepali. Addressing the lack of high-quality training data, we contribute three key datasets: a Nepali and Hindi translation of SQuAD 1.1, a Nepali translation of XQuAD for benchmarking, and a curated Nepali QA dataset derived from Belebele's MCQ data. To mitigate translation-induced answer span loss, we utilize translation-invariant tokens, improving span retention from 50% to 93%, and evaluate translation quality using human assessment and GPT-4, confirming a faithful answer span distribution. We evaluate our models on XQuAD and our curated dataset, demonstrating the effectiveness of fine-tuning multilingual models for Nepali QA. Our best-performing model achieves an exact match (EM) score of 72.99 and an F1 score of 84.13 on XQuAD-Nepali. These results establish a strong baseline for Nepali QA and highlight the impact of utilizing cross-lingual transfer from same language family data. All datasets and code are publicly available, encouraging further advancements in Nepali NLP research.

Keywords — Extractive Question Answering, Low-Resource NLP, Nepali Language, BERT, SQuAD

Introduction

The proliferation of digital information has revolutionized access to knowledge, yet this progress has not been uniform across all languages. Speakers of low-resource languages, such as Nepali, often face significant barriers in accessing information due to the lack of digital tools and resources tailored to their language. This research seeks to address these challenges by developing an extractive question-answering (QA) system for the Nepali language.

In this work, we frame the problem of question answering as a span-based classification task, where the goal is to identify the start and end positions of the answer within the given context. This task formulation aligns with the architecture of modern QA systems based on transformer models, particularly those using the BERT framework. We also focus on evaluating the performance of multilingual models, such as MuRIL and XLM-RoBERTa. These models, pre-trained on a diverse set of languages, offer the advantage of better generalization to low-resource languages as compared to monolingual models. Our approach involves fine-tuning MuRIL [2], a multilingual language model trained on a vast corpus of Indo-Aryan and Indo-Dravadian language data, including Nepali. To overcome the scarcity of Nepali-specific training data, we fine-tune the model using a translated version of the SQuAD dataset. While machine translation offers a solution to the data scarcity problem, we are cognizant of the challenges it introduces, particularly with respect to preserving the integrity and accuracy of answer spans in the translated data. To mitigate these issues, we adopt strategies that enhance the robustness of our system. Additionally, we curate a dedicated Nepali QA dataset and evaluate the system's performance on this dataset, assessing its real-world applicability. This curated dataset, along with the methodologies employed for dataset creation and model fine-tuning, will be made publicly available to facilitate further research and development in Nepali NLP.

The contributions of this research include:

(1) The development of the extractive QA system for Nepali.

(2) An exploration of the effectiveness of fine-tuning multilingual transformer models for extractive Nepali QA. (3) use of a better approach to handle the extraction of answer spans in translated datasets.

* *Corresponding Author*

(4) The creation and release of a Nepali and Hindi translation of SQuAD 1.1, a Nepali translation of XQuAD for benchmarking, and a curated Nepali QA dataset derived from Belebele's MCQ data to facilitate further advancements in Nepali NLP. By addressing the critical need for Nepali language-specific NLP tools, this project aims to empower Nepali speakers with improved access to information and contribute to the advancement of low-resource language processing.

Background and Related Work

The rise of Transformer-based models like BERT [3] has significantly advanced English-language QA systems. These models, pre-trained on extensive English corpora and fine-tuned with large-scale QA datasets such as SQuAD [4], have demonstrated high accuracy. However, this success has not been replicated in low-resource languages such as Nepali. The primary challenge lies in the scarcity of large-scale, high-quality QA datasets.

To address this disparity, researchers have developed multilingual models such as multilingual BERT (mBERT) [5], XLM-RoBERTa (XLM-R) [6], and MuRIL [2]. These models are pre-trained on corpora spanning multiple languages, enabling zero-shot and cross-lingual transfer learning that can benefit low resource languages. In our case especially, models like MuRIL, which is trained specifically focusing on languages of the Indian subcontinent, greatly help the Nepali language due to linguistic similarity and shared features.

Question answer data is another limiting factor in training effective QA models. We will need to use some form of translation to generate sufficient training data, which inevitably risks data loss. Nuutinen et al. [7] proposed a simple yet effective method of preserving spans through translation using translation-invariant tokens, which minimizes this data loss and maintains answer boundary integrity.

Additionally, the study by Kumar et al. [8] proposed to augment QA samples in the target language through translation and transliteration into other languages. This augmented dataset is subsequently employed to finetune a multilingual BERT-based QA model already pre-trained in English. Their experiments using the Google ChAII dataset [9] revealed that fine-tuning the mBERT model with translations from the same language family elevates question-answering performance. However, an inverse effect is observed when employing translations from cross-language families, resulting in degraded performance. In response to these insights, our approach involves translating the SQuAD 1.1 [4] dataset into Hindi, a language that shares considerable linguistic commonality with Nepali. Both languages belong to the Indo-Aryan language family and share grammatical structures, vocabulary, and cultural contexts, making Hindi an ideal language to improve performance for Nepali QA.

Nepali, an underrepresented language in the NLP landscape, faces challenges due to limited linguistic resources. The utilization of Belebele [10], spanning 122 language variants, including the Nepali MCQ dataset, presents a good opportunity for creating an extractive QA dataset in Nepali. Belebele's curated questions, based on short passages from the Flores-200 dataset [11], can serve as an important benchmark for discerning the comprehension levels of the QA model.

Experimental Workflow

Dataset Preparation

The Stanford Question Answering Dataset (SQuAD 1.1) [4] serves as the main dataset for our study. The dataset is divided into 87,599 training examples and 10,570 validation examples. It consists of context and QA pairs. To facilitate the development of extractive QA models in Nepali, we contribute three datasets:

1. *SQuAD-Nepali and SQuAD-Hindi* – We translated the SQuAD 1.1 dataset into Nepali and Hindi, preserving answer spans using translation-invariant tokens.
2. *XQuAD-Nepali* – We translated the XQuAD English dataset into Nepali to serve as an evaluation benchmark.
3. *Belebele-Based Nepali QA Dataset* – We curated a 266 sample Nepali QA dataset derived from the Belebele MCQ dataset, reformatting multiple-choice questions in an extractive QA format.

These datasets, models, and code are made publicly available on Hugging Face and GitHub.

Translation Process

We used Google Translate to translate the data from English to Nepali and Hindi. Each context was translated multiple times for different questions to introduce variability in the translated output. This approach enhances the robustness of our model by incorporating minor variations in translation.

Translating the QA data by individually passing the question, context, and span introduces two issues:

4. *Loss of Answer Spans:* When translating both the context and the answer using the translation model, discrepancies often arise where the answer gets translated differently and cannot be found in the translated context. Out of 87,599 English QA pairs translated into Nepali, only 43,558 pairs retained valid answer spans, resulting in only 50% of the original data.

5. *Change of Answer Length Distribution:* The likelihood of finding the answer span within the translated content was strongly influenced by its length. Short spans tended to remain unchanged or experience minimal alteration during translation, while longer spans were more susceptible to contextual variations. Analysis of the data distribution revealed a clear discrepancy between the original answer spans and their translated counterparts.

The distribution in Figure 1 was obtained by comparing the translated answer spans found in the context and those not found in the context. The spans found after translation were *accepted* in the dataset, whereas the spans not found were *discarded* from the dataset.

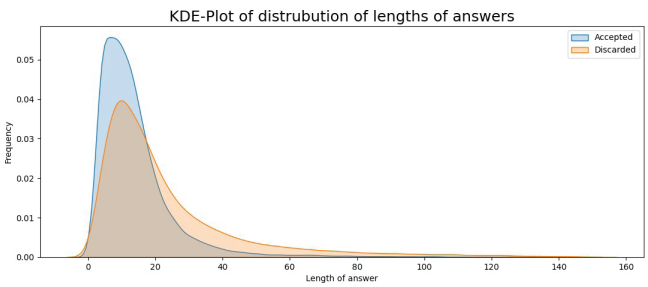

Figure 1: Distribution of Answer Spans with Naive Translation Approach

There was a disproportionate retention of shorter spans compared to longer spans. As the length of the span increases, there is a greater likelihood that a word will be replaced by its synonym and hence not be found. This analysis underscores the need for a more refined approach to identify translated response spans effectively.

Translation-Invariant Tokens

Drawing inspiration from the methodology employed by Nuutinen et al. on Finnish data [7], we adapted their approach for use on Nepali SQuAD data. To mitigate translation challenges, special tokens were encapsulated around the answer span. These tokens were translation-invariant, ensuring their retention during translation. The rationale behind this strategy is to encapsulate the answer span within these invariant markers, enabling us to easily locate and extract the translated answer span post-translation. The process as shown in Figure 2 involved the following steps:

1. *Insertion of Special Tokens:* A special token was placed around the answer span within the context before translation. These tokens served as markers indicating the boundaries of the answer span. For our experiments, the double asterisk (**) was used as the special token, selected arbitrarily.

2. *Translation:* The entire context, including the inserted special tokens, was translated into the target languages: Nepali and Hindi.

3. *Extraction of Translated Answer Span:* Post-translation, the content enclosed within the special tokens was extracted from the translated text. This extracted portion constituted the translated answer span.

4. *Removal of Special Tokens:* Finally, the special tokens were removed from the translated text, leaving behind only the accurately translated answer span.

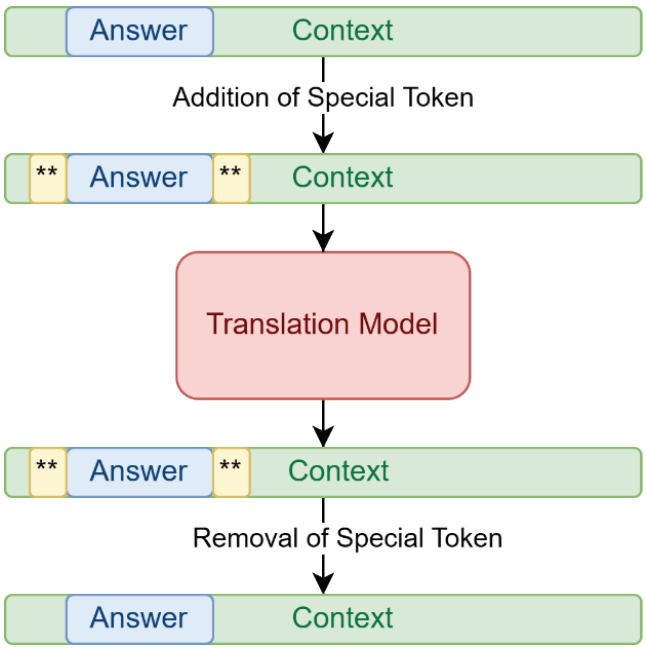

Figure 2: Translation pipeline

Analysis of the Translated Dataset

This approach effectively preserved contextual integrity and significantly enhanced the retention of answer spans to 93%, allowing us to recover 82,036 question-answer pairs out of the 87,599 pairs after translation. Although this approach presents an effective solution, it prompts several inquiries regarding its efficacy and validity.

- *Impact of Special Tokens on Translation Quality:* One concern is whether the presence of the special tokens affects the overall quality of translation. It is essential to investigate whether the insertion of these markers influences the accuracy and fidelity of the translated text.
- *Validity of Encapsulated Answer Span:* Another crucial question revolves around whether the answer span enclosed by the special tokens accurately represents the intended answer. There is a need to verify if the span surrounded by these markers indeed corresponds to the correct answer.

Qualitative Evaluation by Human Assessment

Six evaluators reviewed the translated spans, comparing versions with and without special tokens. The evaluation focused on two primary aspects: assessing the accuracy of the translated spans and discerning any disparities between the versions with and without special characters. The variations spotted between sentences with and without special symbols are:

- Some words underwent different translations, being substituted with synonyms lacking any apparent correlation with the special character.
- Certain tokens, such as नामयोगी (postpositions in Nepali grammar, functioning similarly to English prepositions but appearing after the noun), were either omitted due to the presence of special characters or merged with adjacent words in the answer. An example of this effect is shown in Table 1. Either way, they do not affect the overall coherence and meaning of the sentences, making their impact on translation quality negligible.

Table 1

Effect of Postposition in Nepali Sentence

Language	English	Nepali Translation (Without Special Tokens)	Nepali Translation (With Special Tokens)
Context	These professional qualifications may include the study of **pedagogy**, the science of teaching.	यी व्यावसायिक योग्यताहरूमा शिक्षाशास्त्रको अध्ययन, शिक्षण विज्ञान समावेश हुन सक्छ।	यी व्यावसायिक योग्यताहरूमा **शिक्षाशास्त्र**को अध्ययन, शिक्षण विज्ञान समावेश हुन सक्छ।
Question	What does the study of these professional qualifications focus on?	यी व्यावसायिक योग्यताहरूको अध्ययन केमा केन्द्रित छ?	यी व्यावसायिक योग्यताहरूको अध्ययन केमा केन्द्रित छ?
Answer	pedagogy	शिक्षाशास्त्रको	शिक्षाशास्त्र

With these evaluations, we concluded that while minor variations were observed, the fundamental coherence and fidelity of the translated content remained largely unaffected by the inclusion of special symbols.

Analysis Using Large Language Model

To reduce potential human biases and expand the scope of our analysis, we utilized GPT-4 for sentence compression. Prior research has demonstrated the effectiveness of using large language models for model evaluation. [12] In our study, GPT-4 was provided with a context, a question, and two candidate answers: one generated by a naive translation approach (answer1) and the other by a translation-invariant token method (answer2). The model was then asked to indicate its preference by outputting one of three options: "answer1" if the naive translation was superior, "answer2" if the translation-invariant method was better, or "no change" if both outputs were essentially equivalent.

Upon evaluating a sample of 200 sentence pairs, approximately:

- 60% of the sentence pairs were deemed equivalent, showing no change between the two approaches.
- 18% of the pairs were rated in favor of the naive translation (answer1).
- 22% of the pairs were rated in favor of the translation-invariant token approach (answer2).

While these results provide a quantifiable comparison between the two methods, statistical tests did not reveal significant differences. This suggests that the observed variations may be due to random variation rather than a systematic superiority of one approach over the other. Consequently, it was concluded that there was no significant deviation from the results obtained by this analysis.

Distribution Analysis

Unlike the distribution of our naive method in Figure 1, using this method, the final resulting distribution in Figure 3 is faithful to both shorter and longer answers. This captures the essence of context and answers much better.

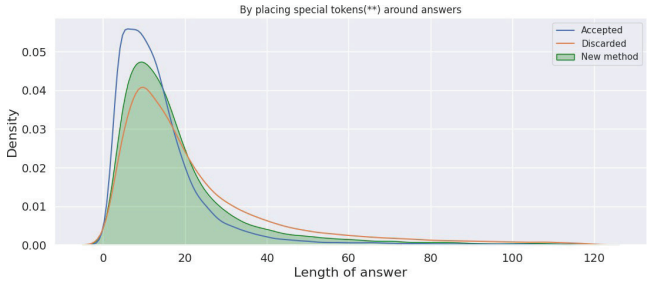

Figure 3: KDE Plot for Result Distribution

Limitations of Translation-invariant Token-based Approach

While translation-invariant tokens substantially improve the answer span retention, preserving 93% of the data, we identified some limitations leading to a 7% loss. On analysis of this, we found the following:

6. Answer Span Loss: In some cases, the translation process failed to preserve the answer span, resulting in its omission.
7. Special Token Misalignment: In other cases, the translation model retained the opening marker but split the closing marker into individual asterisks. This misalignment of special tokens in translated context discards valid answer spans as shown in Table 2.

Question: What did DECnet originally do?

Nepali Translation: DECnet ले शुरुमा के गर्थ्यो ?

Table 2

Example of Token Separation Issue in Translation

Context	Nepali Translation
DECnet is a suite of network protocols created by Digital Equipment Corporation, originally released in 1975 in order to **connect two PDP-11 minicomputers**. ...	DECnet डिजिटल उपकरण निगम द्वारा सिर्जना गरिएको नेटवर्क प्रोटोकलहरूको एक सुइट हो, जुन मूल रूपमा १९७५ मा **दुई PDP-११ मिनी कम्प्युटरहरू* जडान गर्न* जारी गरिएको थियो। ...

We observed a critical issue with span preservation during translation. In this example, the ending token failed to properly capture the verb phrase "जडान गर्न" (connect), which is essential to the answer span's completeness. This misalignment can be attributed to the fundamental structural differences between English and Nepali syntax. English follows a subject-verb-object order, while Nepali follows a subject-object-verb pattern. Consequently, when translating from English to Nepali, the verb phrase shifts position—moving from the beginning of the span in English ("connect") to the end of the span in Nepali ("जडान गर्न"). This sometimes caused the (**) token to be placed incorrectly. To address this, we performed the following methods:

Alternative Marker Strategies: We explored using a single asterisk (*) character as a translation-invariant token. Although a single-character token should be translation invariant and less likely to be split during translation, in practice, our observations showed that passages marked with only one asterisk (*) frequently lost their end tokens. The double asterisk (**) approach significantly reduced this issue, providing more reliable span preservation across translations.

Post-Translation Reassembly: To address the issue of separated closing markers, we implemented a post-processing step that detects and recombines the separated closing markers, ensuring that the answer span boundaries are correctly restored after translation.

Training Procedure

Our starting point is MuRIL-base-cased [2], a multilingual model pre-trained on 17 Indian languages and their transliterated counterparts. Since MuRIL already includes Nepali in its pre-training corpus, we utilize its multilingual capabilities for our downstream question-answering (QA) task. To frame question-answering as a classification task, appropriate preprocessing was conducted. The Sentence Piece tokenizer from MuRIL was used, encoding the input as: [CLS] Question [SEP] Context [SEP], with segment embeddings differentiating question and context text. Two classification heads predict the start and end tokens of the answer span. The final embedding is processed through a linear layer with softmax activation, yielding probability distributions over the token sequence. The model selects the answer span corresponding to the highest probability start-end pair.

For each example, two labels were generated to indicate the answer's start and end tokens. The base model's maximum input length is 512 tokens. Contexts exceeding this limit were split into overlapping segments of 256 tokens. If an answer was not present in a particular segment, the labels will be start = end = 0, marking the [CLS] token as the predicted answer. Offset mappings were used to trace token positions back to the original text.

Our procedure involves:

1. *Initial Fine-Tuning:* We first fine-tune MuRIL on the SQuAD v1.1 dataset, with a training set of 78,839 question-answer pairs and a validation set of 8,760 pairs in English to adapt it from a masked language modeling and next-sentence prediction objective to the extractive QA format.

2. *Language Adaptation*: We then explore two strategies for further fine-tuning on the Nepali and Hindi translations of SQuAD v1.1 so that the model benefits from linguistic similarities and shared structures between languages within the same family, as illustrated in Figure 4.

   - *Sequential Fine-Tuning:* We first fine-tuned on Nepali-translated SQuAD, with a training set of 73,832 question-answer pairs and a validation set of 8,204 pairs. We then further fine-tuned the model on Hindi-translated SQuAD, with a training set of 71,567 question-answer pairs and a validation set of 7,952 pairs. Sequentially fine-tuning on one language at a time can help the model adapt gradually. In many multilingual setups, focusing on a single language before moving to the next can mitigate negative interference [6] across languages. However, this approach can lead to catastrophic forgetting [7] of previously learned language-specific knowledge if the model overfits to the second language's data distribution.
   - *Combined Fine-Tuning:* We fine-tuned on the combined Nepali and Hindi-translated SQUAD data simultaneously, with a training set of 145,399 question-answer pairs and a validation set of 16,156 pairs. Training on mixed data from both languages allows the model to learn shared multilingual representations concurrently. This joint exposure can foster better *cross-lingual transfer* [6], especially when the languages share structural or lexical similarities.

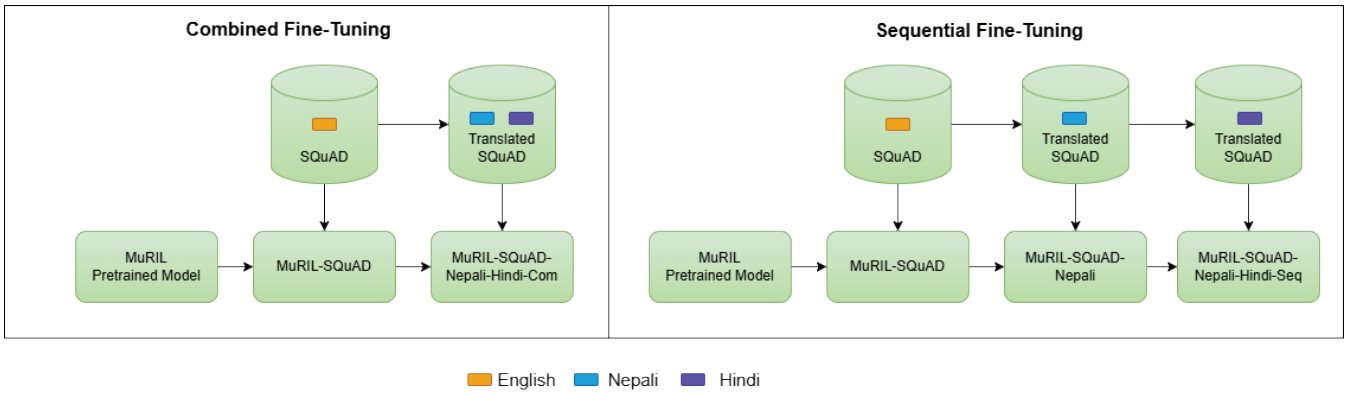

Figure 4: Combined vs. Sequential Fine-Tuning

Empirical Observations

We evaluated both approaches on XQuAD [13], focusing on Hindi (1190 QA pairs) and Nepali (1170 QA pairs). The results are summarized below in Table 3.

Table 3

Performance comparison of fine-tuning approaches on XQuAD Hindi and Nepali datasets.

Model	Dataset	Exact Match	F1
MuRIL-Base-SQuAD-Nepali-Hindi-Seq	XQuAD (Hindi)	59.75	76.18
MuRIL-Base-SQuAD-Nepali-Hindi-Com	**XQuAD (Hindi)**	**60.00**	**76.50**
MuRIL-Base-SQuAD-Nepali-Hindi-Seq	**XQuAD (Nepali)**	**69.74**	**81.03**
MuRIL-Base-SQuAD-Nepali-Hindi-Com	XQuAD (Nepali)	68.72	80.33

Given the minimal performance gap, we opted for the combined fine-tuning strategy in our final pipeline. While sequential fine-tuning can be beneficial for preventing negative interference, it introduces additional complexity and risks catastrophic forgetting if not carefully balanced. Combined fine-tuning simplifies the workflow and allows the model to learn from multiple languages concurrently, which can be advantageous when dealing with resource constraints and time efficiency [6].

RESULTS AND DISCUSSION

The evaluation process involved multiple phases: initial benchmarking on the original SQuAD 1.1 dataset, fine-tuning on translated Nepali and Hindi data, and performance assessment on XQuAD Hindi benchmark, translated XQuAD Nepali and Belebele-based human-annotated Nepali QA dataset. The models were first benchmarked on the English SQuAD dataset. While the MuRIl-Base-SQuAD model performed well in English, it faced challenges with Nepali queries. After fine-tuning models with Nepali and Hindi data, the model showed better adaptation to Nepali QA, as shown in Table 4.

Table 4

Comparison of QA Performance

Context	Pulchowk Campus is one of the five constituent campuses under Institute of Engineering of Tribhuvan University. Situated at Pulchowk of Lalitpur metropolitan city, Pulchowk Campus is the central campus of Institute of Engineering. The campus offers bachelor degree, master degree and doctoral degree programs in? various disciplines. Established in 1972 AD, the campus is second oldest engineering institution of Nepal after Thapathali Campus.	
Question	पुल्चोक क्याम्पसमा कुन-कुन शैक्षिक डिग्री कार्यक्रमहरू उपलब्ध छन्?	
Question Translation	What academic degree programs are offered at Pulchowk Campus?	
Model	**MuRIL-Base-SQuAD**	**MuRIL-Base-SQuAD-Nepali-Hindi**
Answer	doctoral *(Incomplete)*	bachelor degree, master degree and doctoral *(Complete)*

The XQuAD [13] Hindi dataset served as a key benchmark to evaluate model performance. Evaluating existing multilingual models like XLM-Roberta-L and mBERT on this dataset showed that multilingual models, not specifically trained for Nepali question-answering tasks, lagged in performance. The MuRIL-SQuAD-Nepali-Hindi model, trained on both Nepali and Hindi, outperformed other models, achieving an EM of 60.76 and an F1 of 76.59 for the base model and an EM of 64.62 and an F1 of 80.14 for the large model. The model trained on both Hindi and Nepali outperformed the model trained on Nepali data only, which highlights the model's cross-lingual transfer capability and its ability to leverage linguistic similarities between Hindi and Nepali.

We also translated the XQuAD English data into Nepali using translation-invariant tokens and obtained 1170 QA pairs. Our MuRIL-based models demonstrated superior proficiency in handling Nepali queries and achieved an EM of 72.99 and an F1 of 84.13.

We created 266 Nepali QA dataset using Belebele's MCQ dataset and Haystack annotation tool[5] and used it to measure the performance of our models against human-annotated dataset. Although the performance on this dataset was lower—achieving an EM of 21.80 and an F1 score of 50.51 for the MuRIL-Large-SQuAD-Nepali-Hindi model—the drop in metrics is attributable to a change in distribution of the standard SQuAD dataset and the human-annotated dataset. Despite this, the model serves as a good starting point for further research work in this domain.

Table 5

Performance on XQuAD (Hindi, Nepali) and Nepali Curated Dataset

Model	Exact Match	F1
XQuAD (Hindi, 1190 QA pairs)		
XLM-Roberta-L	53.60	70.3
mBERT	46.0	59.2
MuRIL-Base-SQuAD (Ours)	47.0	63.85
MuRIL-Base-SQuAD-Nepali (Ours)	57.22	73.22
MuRIL-Base-SQuAD-Nepali-Hindi (Ours)	**60.76**	**76.59**
MuRIL-Large-SQuAD (Ours)	62.19	78.30
MuRIL-Large-SQuAD-Nepali-Hindi (Ours)	**64.62**	**80.14**
XQuAD (Nepali, 1170 QA pairs)		
MuRIL-Base-SQuAD-Nepali-Hindi (Ours)	70.09	81.21
MuRIL-Large-SQuAD-Nepali-Hindi (Ours)	72.99	84.13
Nepali Curated Dataset (266 QA pairs)		
MuRIL-Base-SQuAD-Nepali-Hindi	19.17	46.41
MuRIL-Large-SQuAD-Nepali-Hindi (Ours)	21.80	50.51

Table 6

Hyperparameters Used for Training

Hyperparameter	MuRIL-Base-SQuAD-Nepali-Hindi	MuRIL-Large-SQuAD-Nepali-Hindi
Learning Rate	2e-05	3e-05
Number of Epochs	4	2
Weight Decay	0.01	0.01
Batch Size	64	32
Seed	42	42
Optimizer	Adam with betas=(0.9,0.999) and epsilon=1e-08	Adam with betas=(0.9,0.999) and epsilon=1e-08
Lr Schedular Type	Linear	Linear
Mixed Precision Training	Native AMP	Native AMP
Max Sequence Length	512	384
Stride	256	128
N-Best Predictions	20	20
Max Answer Length	30	30

Conclusion

In conclusion, this paper presents an Extractive Nepali Question Answering System to address the lack of NLP tools and high-quality QA datasets for Nepali language. We created a Nepali and Hindi translation of SQuAD 1.1 and utilized the cross-lingual transfer between the same language family to improve the model performance. The use of translation-invariant tokens proved crucial in mitigating data loss during translation and preserving answer span integrity. This approach retained 93% of the original dataset, a significant improvement over the 50% retention rate of a naive translation method. Both human evaluation and GPT-4 analysis confirmed the fidelity of the translated data and the effectiveness of the translation-invariant token approach. We also created a Nepali translation of XQuAD for benchmarking and curated Nepali QA dataset from Belebele's MCQ data. Our evaluation on the XQuAD-Nepali dataset demonstrated the effectiveness of fine-tuning multilingual models for Nepali QA. The best-performing model, MuRIL-Large-SQuAD-Nepali-Hindi, achieved an Exact Match (EM) score of 72.99 and an F1 score of 84.13, establishing a strong baseline for future Nepali QA research. The datasets and methodologies presented in this work are made publicly available to encourage further research and development in Nepali NLP.

Acknowledgement

We would like to express our sincere gratitude towards the faculty members of the Department of Electronics and Computer Engineering, IOE Pulchowk Campus, for their constant support and guidance while conducting this research.

References

[1] Encyclopedia Britannica, "Nepali language," https://www.britannica.com/topic/Nepali-language, 2023.

[2] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, and P. Talukdar, "Muril: Multilingual representations for indian languages," 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[4] 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

[5] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" 2019. [Online]. Available: https://arxiv.org/abs/1906.01502

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2020.

[7] E. Nuutinen, I. Rastas, and F. Ginter, "Finnish squad: A simple approach to machine translation of span annotations," 08 2023.

[8] G. K. Kumar, A. S. Gehlot, S. S. Mullappilly, and K. Nandakumar, "Mucot: Multilingual contrastive training for question-answering in low-resource languages," 2022.

[9] M. Sabane, O. Litake, and A. Chadha, "Breaking language barriers: A question answering dataset for hindi and marathi," 08 2023.

[10] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettle moyer, and M. Khabsa, "The belebele benchmark: a parallel reading comprehension dataset in 122 language variants," 2023.

[11] M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard et al., "No language left behind: Scaling human-centered machine translation," arXiv preprint arXiv:2207.04672, 2022.

[12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, L. Zhuohan, D. Li, E. Xing, H. Zhang, J. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 06 2023.

[13] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," CoRR, vol. abs/1910.11856, 2019. [Online]. Available: http://arxiv.org/abs/1910.11856