Proceedings of 15th IOE Graduate Conference
Peer Reviewed
Year: 2024 Month: May Volume: 15
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Nepali Question Answering System from Multilingual BERT Model and Monolingual BERT Model

Upanshu Thapa [a], Suresh Timilsina [b], Hom Nath Tiwari [c], Mukunda Upadhyay [d]

[a, b, c, d] *Department of Electronics and Computer Engineering, IOE, Pashchimanchal Campus, Tribhuvan University, Nepal*
[a] upanshuthapa21@gmail.com, [b] timilsinasurace@gmail.com, [c] homnath@wrc.edu.np, [d] updmuku24@gmail.com

Abstract

A question answering system typically employs methods from Natural Language Processing, machine learning, and knowledge representation to grasp the intent of questions and deliver suitable responses from a related context in a dataset or knowledge base. This paper focuses on developing a Nepali language question answering system by incorporating advanced techniques utilizing both multilingual BERT (Bidirectional Encoder Representations from Transformers) and monolingual BERT models. The research is divided into three phases: preparation of Nepali language datasets, using them to train multilingual BERT and monolingual BERT models, and finally fine-tuning with Nepali question-answer datasets. The primary goal is to explore the potential implementation of these language models in addressing question answering in the Nepali language. To facilitate this investigation, a specialized Nepali question-answering dataset is created by translating and standardizing SQuAD datasets. The study employs both multilingual BERT and monolingual BERT models to train the Nepali Question Answering System. Multilingual BERT, capable of handling multiple languages simultaneously, is utilized to tap into the broader linguistic context shared among different languages, while monolingual BERT is tailored exclusively for Nepali, providing a focused and language-specific approach to the system's development. The research emphasizes the importance of leveraging larger datasets in Nepali, as they significantly contribute to the models' training efficacy. By combining the strengths of multilingual and monolingual BERT models and fine-tuning them on a Nepali question-answering dataset, it is found that the multilingual BERT model, mBERT, performs better than monolingual BERT model, NepBERTa with F1 score of 0.75 compared to 0.66.

Keywords

NLP,Multilingual BERT, Monolingual BERT, Nepali Question-answering

1. Introduction

Natural Language Processing (NLP) can be extensively utilised for tackling question answering tasks. Noticeable efforts have been made in high-level language research to address reading comprehension-based question answering tasks. where answers are extracted from related passages in response to provided input questions presented in natural language. Nepali is the official language of Nepal; however, it is also a native language for. people in some parts of India, Bhutan and Myanmar. Myanmar. It is based on Devanagari scripts. Nepal is a low-resource language, but it is morphologically rich and does not have sufficient datasets like English for question-answering tasks. English language datasets are available like Stanford Question Answering Dataset (SQuAD 2.0)[1] which can solve English question answering tasks with high accuracy. Techniques like transformer-based models, such as BERT, have shown significant advancements in this domain[2]. These systems utilise pre-trained language models that take into account the entire context of a passage or document, capturing bidirectional relationships between words. By considering the surrounding words and sentences, context-aware QA systems can better comprehend nuances, references, and co-references within the text, leading to more accurate answers. BERT is a powerful natural language processing model introduced by Google in 2018 [2]. The transformer model relies on self-attention mechanisms to process input data in parallel, making it highly efficient for sequence-to-sequence tasks. It considers the context from both the left and right sides of each word in a sentence. This bidirectionality enables BERT to understand the relationships between words in a more nuanced way. BERT is pre-trained on massive amounts of text data to learn contextualized representations of words.

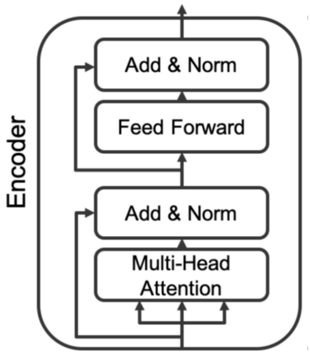


Figure 1: Architecture of encoder layer used in BERT

After pre-training, BERT can be fine-tuned on specific downstream tasks such as question answering, sentiment analysis, or named entity recognition[2]. Fine-tuning involves training the model on task-specific labeled datasets to adapt its knowledge to the target application. One of BERT's key advantages is transfer learning. Pre-trained on a large corpus, BERT can be fine-tuned on smaller, task-specific datasets, even with limited labeled examples. This enables the model to generalize well to various NLP tasks. BERT's bidirectional contextual embedding and pre-training on massive corpora have contributed to its breakthrough performance on various NLP tasks, making it a cornerstone in contemporary natural language processing research. Question answering using BERT involves fine-tuning the pre-trained BERT model on a specific QA dataset. BERT, being a transformer-based language model that captures contextual relationships between words bidirectionally, is highly effective for understanding context in natural language. The fine-tuned BERT model, specific to question answering, can then provide accurate and contextually relevant answers by considering the bidirectional context of the input text. Fine-tuning is essential to adapt the model to the specific nuances and characteristics of the QA task. The structural diagram of the encoder layer used in BERT is shown in Figure 1.

We delve into the challenges of accurately translating English text into Nepali scripts to generate top-notch datasets. Leveraging Google Translate, often regarded as a proficient tool for language translation, we aim to streamline dataset creation for low-resource languages. However, Google's translations may not always yield precise semantic results. To address this, we meticulously reviewed translations manually, re-translating where necessary, and discarded poorly translated entries [3]. Subsequently, we trained monolingual and multilingual transformer-based language models tailored for Nepali. Their performance was evaluated by fine-tuning them using the Nepali reading comprehension dataset.

This research is organized into five sections. Firstly, we start with a basic introduction followed by Section 2 related works, which narrows down our motivation and research gap. The research issue concerns the unavailability of datasets for Nepali question answering and the need to tackle its implementation. Section 3 explains data preprocessing and dataset creation, and also provides a limelight on the entire proposed methodology along with our problem description. The datasets for Nepali question answering were created following the standard SQuAD format and underwent fine-tuning processes utilizing both language-specific and multilingual models. Section 4 reports the results and discussion of the implemented model. We conclude in Section 5, highlighting some future works too.

2. Related Works

Transformer-based models have been implemented in various Natural Language Processing tasks such as question answering, machine translation, text recognition, text-summarization, news classification and part-of-speech tagging. Numerous transformer models have been created specifically for the English language, including Bidirectional Encoder Representations from Transformers (BERT) [2] and Generative Pretrained Transformer (GPT) [4]. One of the popular multilingual models is multilingual BERT (mBERT) [5]. NepBERTa is a BERT-based Natural Language Understanding (NLU) model that has been trained on the largest monolingual Nepali corpus to date [6]. It assessed the effectiveness of NepBERTa across various Nepali-specific Natural Language Processing tasks, such as Named-Entity Recognition, Part-of-Speech Tagging, and Sequence Pair Similarity. It indicates the task of Nepali-Question answering system as a research gap to be fulfilled which can be implemented in digital format for business, study and entertainment purpose.

Wagle et al. performed a comparative analysis of news classification in the Nepali language [7]. The comparative study was conducted using various models such as LSTM, BiLSTM, and transformer models. In this study, the Transformer model outperformed other models, leading to the selection of the transformer model BERT for Nepali language tasks. Tripathi conducted Sentiment Analysis of Nepali COVID19 Tweets [8]. This study was conducted on Nepali language tweets and applied various models such as SVM and LSTM. The application of Natural Language Processing in sentiment analysis in the Nepali language was crucial. Devlin et al. introduced the language model called BERT [2], which is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

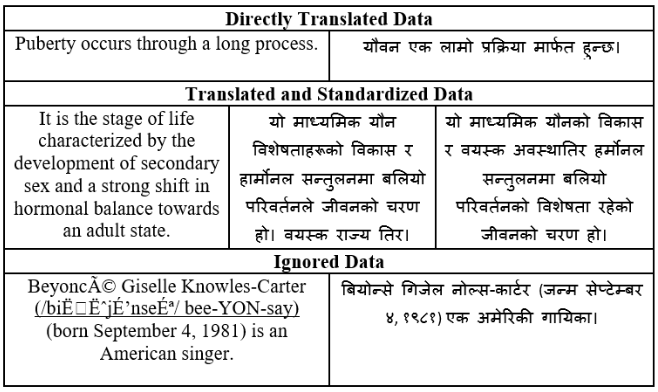
Sarkar et al. perform the sentiment analysis for documents [9]. It provides a method to build a sentiment classifier for a language in which we have no labelled sentiment data. This method addresses the limited availability of labeled data and offer a fresh approach to accomplishing downstream tasks in regional languages. Chen et al. rank a list of candidate passages in multiple languages given a query [10]. The task of Cross-lingual Paraphrase Identification aims to identify whether two sentences, which may be from different languages, have the same meaning. This research gave the idea to apply the multilingual datasets in pre-trained BERT model. Vukadin et al. uses mBERT to extract useful information from unstructured CV documents[11]. The model is designed to discern relevant sections of documents, such as personal information, education, and employment, along with specific details at a lower hierarchy level, including names, addresses, roles, and skill competences. The methodology employs the transformer architecture, utilizing the BERT language model in its multilingual implementation for the encoder part.

Mutabazi et al. performed the question answering system in medical datasets [12]. Different types of deep learning techniques were used to perform question answering in different medical datasets. This paper provides different evaluation metrics for question answering like accuracy, F1 score, MAP and MRR. If multiple correct answers are possible, F1 could be more appropriate. Mean Reciprocal Rank is used to evaluate when the ranked list of answers are provided. Trang et al. uses different BERT models to question answer low resource Vietnamese Language [13]. Some multilingual BERT fine-tuned models support the Vietnamese language. The language-specific models still outperform multilingual ones in QA task. Kulkarni et al. uses different BERT models to perform question answering in low resource Marathi language [14]. The SQuAD v1.1 dataset used in this research served as the foundational dataset, offering researchers ample data to formulate solutions for addressing the question-answering problem. A potential initial step in constructing a deep learning-based question-answering system for Marathi involves translating existing English datasets into the Marathi language.

3. Methodology

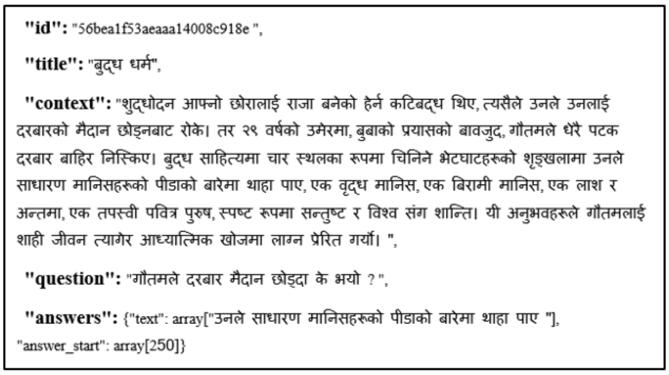
3.1 Dataset Creation

Applying techniques such as translating, standardizing, and filtering the standard English language SQuAD2.0 dataset, Nepali datasets were created. Google Translate was utilized for translating the data. Some datasets were transliterated with the assistance of secondary level Nepali language teacher due to the absence of proper translations, aiming to standardize our datasets, while non-translatable data were disregarded, as depicted in Figure 2. The final translated dataset, post-filtration, comprises 825 question-answer and passage pairs for training. The complete dataset includes 78 unique contexts, 814 unique questions, and 768 unique answers. Figure 3 illustrates the organization of a singular context along with its corresponding question answers within the Nepali dataset. Each passage is associated with multiple questions, primarily comprising factoid queries, and concise answers are provided for such questions. It's noteworthy that all questions in this dataset adhere to the answerable question format, aligning with SQuAD2.0.



<table>
<tr><th colspan="6">Directly Translated Data</th></tr>
<tr><td colspan="3">Puberty occurs through a long process.</td><td colspan="3">यौवन एक लामो प्रक्रिया मार्फत हुन्छ।</td></tr>
<tr><th colspan="6">Translated and Standardized Data</th></tr>
<tr><td colspan="2">It is the stage of life characterized by the development of secondary sex and a strong shift in hormonal balance towards an adult state.</td><td colspan="2">यो माध्यमिक यौन विशेषताहरूको विकास र हार्मोनल सन्तुलनमा बलियो परिवर्तनले जीवनको चरण हो। वयस्क राज्य तिर।</td><td colspan="2">यो माध्यमिक यौनको विकास र वयस्क अवस्थातिर हर्मोनल सन्तुलनमा बलियो परिवर्तनको विशेषता रहेको जीवनको चरण हो।</td></tr>
<tr><th colspan="6">Ignored Data</th></tr>
<tr><td colspan="3">BeyoncÃ© Giselle Knowles-Carter (/biË□Ë^jÉ'nseÉª/ bee-YON-say) (born September 4, 1981) is an American singer.</td><td colspan="3">बियोन्से गिजेल नोल्स-कार्टर (जन्म सेप्टेम्बर ४, १९८१) एक अमेरिकी गायिका।</td></tr>
</table>

Figure 2: Preparation of datasets



"id": "56bea1f53aeaaa14008c918e ",

"title": "बुद्ध धर्म",

"context": "शुद्धोदन आफ्नो छोरालाई राजा बनेको हेर्न कटिबद्ध थिए, त्यसैले उनले उनलाई दरबारको मैदान छोड्नबाट रोके। तर २९ वर्षको उमेरमा, बुबाको प्रयासको बावजुद, गौतमले धेरै पटक दरबार बाहिर निस्किए। बुद्ध साहित्यमा चार स्थलका रूपमा चिनिने भेटघाटहरूको शृङ्खलामा उनले साधारण मानिसहरूको पीडाको बारेमा थाहा पाए, एक वृद्ध मानिस, एक बिरामी मानिस, एक लाश र अन्तमा, एक तपस्वी पवित्र पुरुष, स्पष्ट रूपमा सन्तुष्ट र विश्व संग शान्ति। यी अनुभवहरूले गौतमलाई शाही जीवन त्यागेर आध्यात्मिक खोजमा लाग्न प्रेरित गर्यो। ",

"question": "गौतमले दरबार मैदान छोड्दा के भयो ? ",

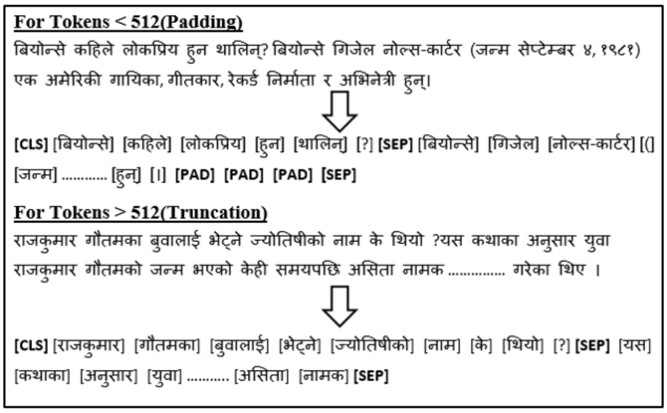
"answers": {"text": array["उनले साधारण मानिसहरूको पीडाको बारेमा थाहा पाए "], "answer_start": array[250]}

Figure 3: Datasets in SQuAD2.0 format

3.2 Data Preprocessing

The preprocessing steps, which involve tokenization, padding, truncation, positional embedding, and word embedding, laid the foundation for the effective utilization of BERT models in question-answering tasks by providing the model with structured and properly formatted input data.

Tokenization and Padding: Tokenization broke the text into individual tokens or sub-word units. These tokens included words, sub-words, and special tokens like [CLS] (classification), [SEP] (separator), and [PAD] (padding). Padding and truncation, as shown in Figure 4, ensured that the input was padded to the maximum length allowed by BERT and truncated if it exceeded that length. Padding tokens were added to the context whenever the token length was less than 512, while truncation ensured the token length remained within 512 tokens. If there were more than 512 tokens, excess tokens would be removed.



For Tokens < 512(Padding)

बियोन्से कहिले लोकप्रिय हुन थालिन्? बियोन्से गिजेल नोल्स-कार्टर (जन्म सेप्टेम्बर ४, १९८१) एक अमेरिकी गायिका, गीतकार, रेकर्ड निर्माता र अभिनेत्री हुन्।

⇩

[CLS] [बियोन्से] [कहिले] [लोकप्रिय] [हुन] [थालिन्] [?] **[SEP]** [बियोन्से] [गिजेल] [नोल्स-कार्टर] [(] [जन्म] .... [हुन्] [।] **[PAD]** **[PAD]** **[PAD]** **[SEP]**

For Tokens > 512(Truncation)

राजकुमार गौतमका बुवालाई भेट्ने ज्योतिषीको नाम के थियो ?यस कथाका अनुसार युवा राजकुमार गौतमको जन्म भएको केही समयपछि असिता नामक .... गरेका थिए ।

⇩

[CLS] [राजकुमार] [गौतमका] [बुवालाई] [भेट्ने] [ज्योतिषीको] [नाम] [के] [थियो] [?] **[SEP]** [यस] [कथाका] [अनुसार] [युवा] .... [असिता] [नामक] **[SEP]**

Figure 4: Tokenization and padding

Positional and Segment Embedding: "Positional embedding refers to supplementary embeddings that encode information about the position or sequence order of words within a given sequence. As BERT processes input tokens concurrently, it lacks inherent information about the order of tokens. It becomes essential for the model to grasp the sequential structure of the input [14]. The segment ID was employed to generate segment embeddings, aiding the model in comprehending which words pertained to the question and which pertained to the context. The segment embedding vector was then combined with the token embedding vector to produce an integrated representation for each term in the input sequence."

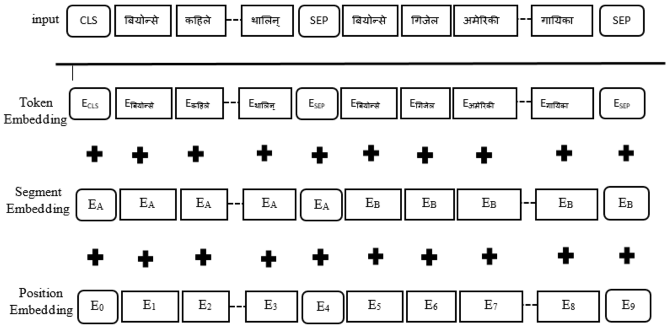

input CLS बियोन्से कहिले थालिन् SEP बियोन्से गिजेल अमेरिकी गायिका SEP
Token Embedding E_{CLS} $E_{बियोन्से}$ $E_{कहिले}$ $E_{थालिन्}$ E_{SEP} $E_{बियोन्से}$ $E_{गिजेल}$ $E_{अमेरिकी}$ $E_{गायिका}$ E_{SEP}
Segment Embedding E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embedding E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9

Figure 5: Positional and segment embedding

3.3 Model

The development of a Nepali question answering system commenced with the creation of a specialized dataset tailored specifically for Nepali question answering. This dataset encompassed a diverse array of Nepali questions and their corresponding answers, spanning various contexts and topics. Various language transformer models pre-trained on large corpora were then explored and selected. After the selection process, these models underwent fine-tuning using the Nepali question answering dataset, enabling them to grasp the intricacies of the Nepali language and optimize their performance. This process of creating Nepali question answer datasets, training them in the best-suited models (mBERT and NepBERTa), and fine-tuning our work to obtain relevant answers to questions from the context is illustrated in Figure 6.

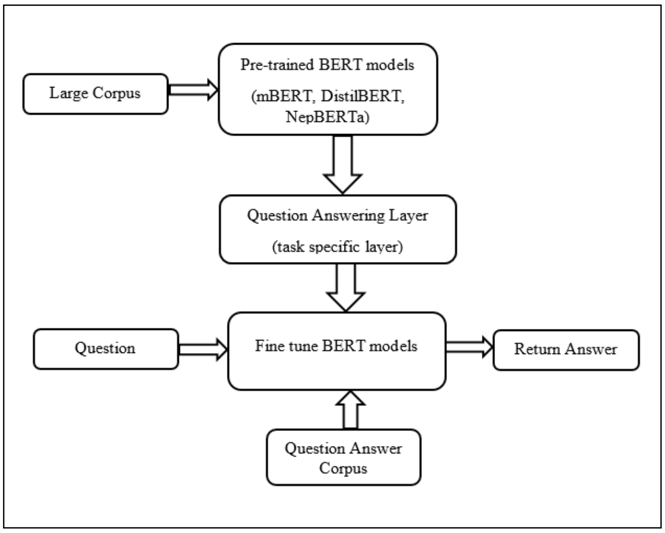


Figure 6: Workflow of BERT question answering

BERT: This model comprehends word relationships by pre-training on extensive unlabeled text and fine-tuning for specific tasks. Unlike linear models, BERT considers both directions simultaneously, grasping each word's context. Through tasks such as 'Masked Language Model' and 'Next Sentence Prediction', BERT captures linguistic nuances. BERT's transformer-based architecture captures long-range dependencies effectively, generating high-quality contextual word representations for various NLP tasks. It utilizes a transformer-based architecture, allowing dynamic weighing of word importance.

Multilingual BERT: This model learns from multiple languages simultaneously, facilitating cross-lingual tasks without separate models. Though proficiency may vary, it offers language-general capabilities efficiently. BERT multilingual is trained on 104 languages, including Nepali, using the Wikipedia dataset and a masked language modeling and next sentence prediction approach. Nepali was one of the languages used for training mBERT transformer models [2]. We utilized 'bert base multilingual cased' version of mBERT for our experiments.

Monolingual BERT: This model is trained specifically on text data from a single language. For example, there are BERT models trained solely on Hindi, Nepali, or other individual languages. Because they are trained on data from a single language, monolingual BERT models may excel in understanding the nuances, syntax, and semantics of that particular language. NepBERTa fine-tuned on a Nepali monolingual corpus consisting of 768 embedding sizes and 12 attention heads, with 110 million parameters was considered to accomplish Nepali question answer task [6].

3.4 Question answering using BERT

In the task of Question Answering, BERT processes both the question and context as one sequence. We merge the question and context using special tokens, namely CLS and SEP, to form the input sequence for the model. The model outputs a text span that comprises the answer, requiring BERT to predict both the start and end tokens of the answer.

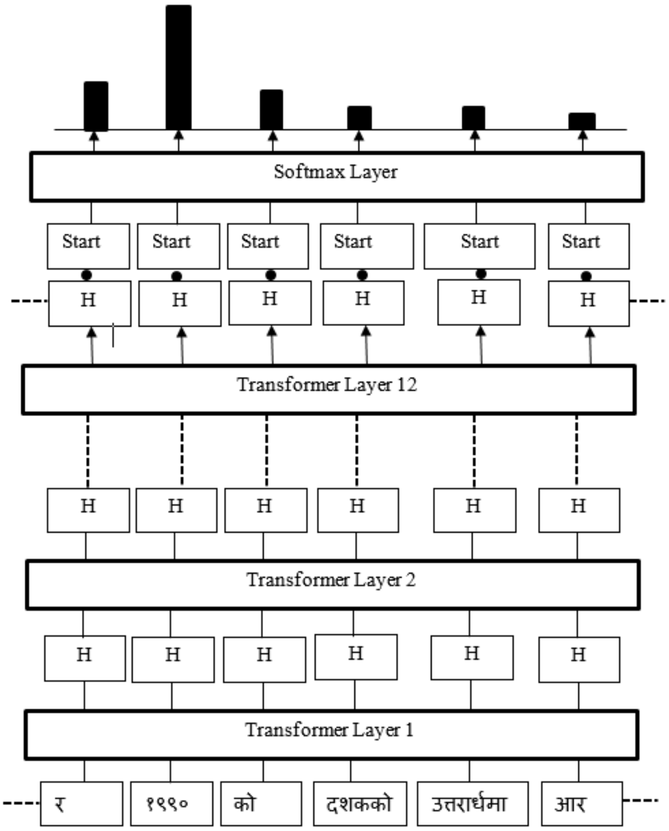


Figure 7: Fine tuning of BERT to generate start token

After the input tokens have been passed through layers of transformers, the model generates word embeddings and introduces a start vector. The probability of each word being the start word is calculated by taking a dot product between the final embedding of the word and the start vector, and SoftMax is applied over all the words. The word with the highest probability value is considered the start of the answer, as shown in Figure 7. The same process is repeated to find the end word of an answer. Once the start and end words have been found, a full answer text can be generated from the context by grabbing the start and end words.

The softmax function defined as follows:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \tag{1}$$

where x_i represents the input logits for token i, and the summation in the denominator is over all tokens in the vocabulary, including the start and end tokens [15].

The softmax function played a crucial role in converting raw model outputs into probability distributions. The softmax function was applied to the start and end logits separately to obtain probabilities for each token in the input sequence, determining whether it could be the start or end of the answer span.

4. Result and Analysis

4.1 Experimental Setup

Fine-tuning was conducted on a Nepali dataset consisting of 825 question-answer and passage pairs using both mBERT and NepBERTa. The dataset encompasses 78 unique contexts, 814 unique questions, and 768 unique answers. The outcome of the fine-tuned models provides probable starting and ending indices for the predicted answers across all models. We performed an 80 percent to 20 percent train-test split in our dataset. The training duration for each model spanned 18 epochs with a batch size set to 8, utilizing GPU processing and a sequence length of 512. On average, the training process for 18 epochs took approximately half to one hour for each model. Cross-Entropy Loss is a metric that quantifies the dissimilarity between the predicted probability distribution assigned to potential answer spans by the model and the actual distribution observed in the ground truth data. The loss and accuracy during the training and validation phases in mBERT are depicted in Figure 8 and Figure 10, respectively, while for NepBERTa, they are shown in Figure 9 and Figure 11. As the figures indicate, the learning curve is better for mBERT, exhibiting lower loss and higher accuracy. Early stopping was applied after there was no improvement from the 17th epoch.

4.2 Evaluation Metrics

When presented with a question and passage, the model aims to predict the most accurate answer by determining the starting and ending indices within the passage. Predicted sequences may contain words not present in the ground truth answers, and they could exhibit semantic similarities to the words in the ground truth. Evaluation metrics, including the Exact Match (EM) score, F1 score, serve as benchmarks for assessing the performance of question-answering models.

The F1 score formula with start and end tokens is typically used in tasks such as question answering and text generation. It can be represented as:

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

where precision and recall are calculated based on the generated text sequence, considering the start and end tokens [16].

BLEU (Bilingual Evaluation Understudy) is a metric employed for assessing natural language generation tasks, such as question answering. The BLEU score evaluates the quality of the generated output by considering the precision of n-grams, which are consecutive sequences of n items, in relation to reference outputs. Precision is defined as the ratio of overlapping n-grams to the total number of n-grams in the predicted output. A Brevity Penalty is incorporated to penalize systems that produce shorter outputs, comparing the length of the predicted output to the reference output. The BLEU score is obtained by multiplying the geometric mean derived from the n-gram precision by the brevity penalty.

The BLEU score formula is defined as:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) \tag{3}$$

where, BP is the brevity penalty, w_n is the weight for n-grams, and p_n is the precision for n-grams [16].

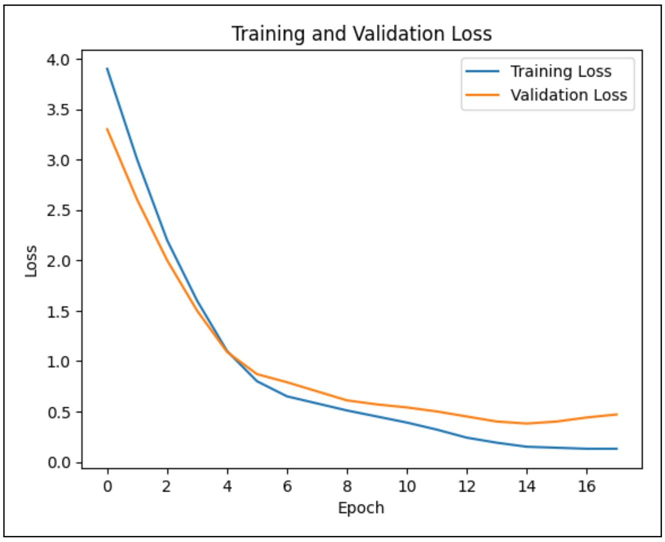


Figure 8: Loss curve in mBERT

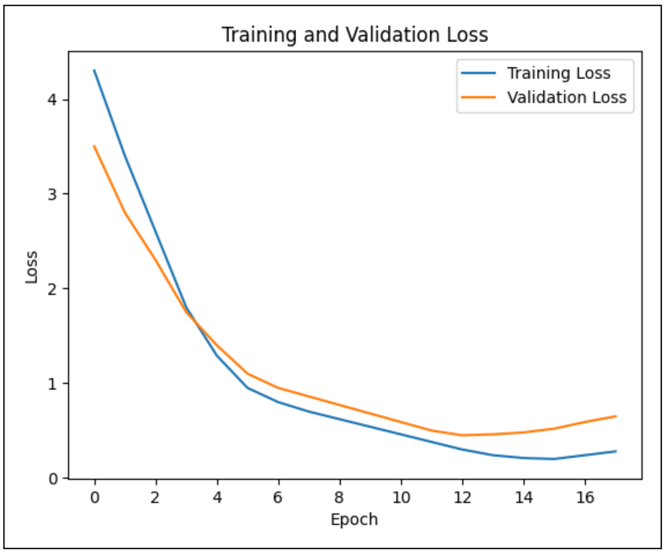


Figure 9: Loss curve in NepBERTa

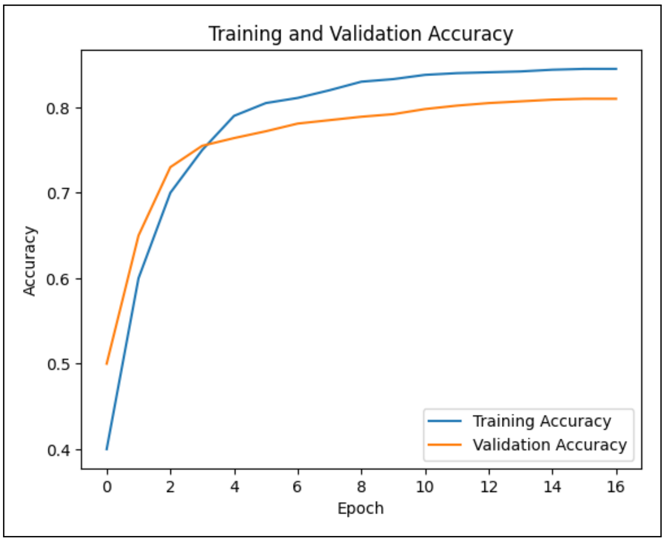


Figure 10: Accuracy curve in mBERT

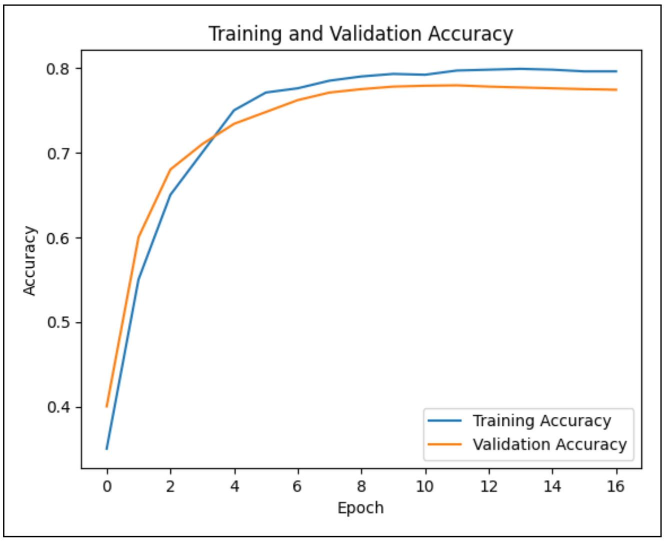


Figure 11: Accuracy curve in NepBERTa

4.3 Results

A fine-tuning was conducted on BERT base multilingual cased and NepBERTa using the Nepali Question Answering datasets, employing the hyperparameters mentioned in the experimental setup. Different evaluation metrics discussed in the previous section were calculated for each model, and the results of each preprocessing step were also observed carefully to see their impact on the final output. The F1 score and BLEU score obtained in mBERT and NepBERTa are listed in Table 1.

Among the two BERT models applied to Nepali question answering datasets, the multilingual BERT model, mBERT, provides the best results. It achieves an F1 score of 0.75, indicating that the answers generated by mBERT are understandable and reasonably accurate. The BLEU scores for both models are in the early 60 percent range, suggesting a moderate level of understanding of the Nepali language. Although the Nepali BERT model, NepBERTa, also performs reasonably well, it lags behind mBERT in terms of accuracy. The results are influenced by the tokenization process performed by these models, where mBERT once again outperforms the other models.

An F1 score of 0.75 in the Nepali question answering dataset can be considered a respectable performance, as it indicates that the model is able to accurately identify relevant information and provide meaningful answers to the questions asked.

Table 1: Results comparing different models

Model	Types of BERT	F1 score	BLEU score
mBERT	multilingual	0.75	0.61
NepBERTa	monolingual	0.66	0.60

5. Conclusion and Future Work

In this research, the Nepali question answering task was conducted using state-of-the-art BERT models. The development of a Nepali question answering system using both multilingual BERT model and monolingual BERT model has been a significant step toward natural language understanding in the Nepali language. Nepali question answering datasets were prepared in the standard SQuAD format using annotation tools. These datasets were trained and validated using different multilingual BERT models, namely mBERT, and the monolingual BERT model, namely NepBERTa. Multilingual models were selected based on the inclusion of the Nepali language within their pre-trained data corpus. The multilingual BERT model mBERT outperforms the monolingual BERT model NepBERTa in every evaluation metric we used. Therefore, mBERT is suggested for performing the Nepali question answering task over NepBERTa.

By adopting larger datasets, we can further improve the performance of pre-trained models. The monolingual BERT model can be developed and implemented in Nepali question answering data to be more accurate and efficient.

References

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Rabin Budhathoki and Suresh Timilsina. Image captioning in nepali using cnn and transformer decoder. *Journal of Engineering and Sciences*, 2(1):41–48, 2023.doi:10.3126/jes2.v2i1.60391.

[4] Aarushi Phade and Yashodhara Haribhakta. Question answering system for low resource language using transfer learning. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, pages 1–6. IEEE, 2021.

[5] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

[6] Sulav Timilsina, Milan Gautam, and Binod Bhattarai. Nepberta: Nepali language model trained in a large

corpus. In *Proceedings of the 2nd conference of the Asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing*. Association for Computational Linguistics (ACL), 2022.

[7] Saroj Sharma Wagle and Sharan Thapa. Comparative analysis of nepali news classification using lstm, bi-lstm and transformer model. 2021.

[8] Milan Tripathi. Sentiment analysis of nepali covid19 tweets using nb svm and lstm. *Journal of Artificial Intelligence*, 3(03):151–168, 2021.

[9] Anindya Sarkar, Sujeeth Reddy, and Raghu Sesha Iyengar. Zero-shot multilingual sentiment analysis using hierarchical attentive network and bert. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, pages 49–56, 2019.

[10] Dongmei Chen, Sheng Zhang, Xin Zhang, and Kaijing Yang. Cross-lingual passage re-ranking with alignment augmented multilingual bert. *IEEE Access*, 8:213232–213243, 2020.

[11] Davor Vukadin, Adrian Satja Kurdija, Goran Delač, and Marin Šilić. Information extraction from free-form cv documents in multiple languages. *IEEE access*, 9:84559–84575, 2021.

[12] Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456, 2021.

[13] Nguyen Thi Mai Trang and Maxim Shcherbakov. Vietnamese question answering system f rom multilingual bert models to monolingual bert model. In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 201–206. IEEE, 2020.

[14] Dhiraj Amin, Sharvari Govilkar, and Sagar Kulkarni. Question answering using deep learning in low resource indian language marathi. *arXiv preprint arXiv:2309.15779*, 2023.

[15] Meiqi Wang, Siyuan Lu, Danyang Zhu, Jun Lin, and Zhongfeng Wang. A high-speed and low-complexity architecture for softmax function in deep learning. In *2018 IEEE asia pacific conference on circuits and systems (APCCAS)*, pages 223–226. IEEE, 2018.

[16] Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.