Proceedings of 14th IOE Graduate Conference
Peer Reviewed
Year: 2023 Month: December Volume: 14
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Next Nepali Word Prediction Using LSTM and BiLSTM

Bikram Shah [a], Manoj K. Guragai [b]

[a, b] *Department of Electronics and Computer Engineering, Purwanchal Campus, IOE, Tribhuvan University, Nepal*
[a]bikramshah213@gmail.com, [b]mkguragai@gmail.com

Abstract
Language modelling, or the act of predicting the next word in a sentence, is a branch of natural language processing (NLP). The main aim is to predict the next word in the sequence. It plays a vital role in various applications such as text generation, machine translation, auto-completion, etc. This paper focuses on advancing the performance of next-word prediction models by utilizing LSTM and BiLSTM models in the Nepali Language. Traditional n-gram-based language models often fail to capture the intricate contextual relationships between words, leading to suboptimal predictive performance. To address this limitation, a novel approach that uses LSTM and BiLSTM models have been developed in this work that predicts the next Nepali words in the sentence. This paper presents a comprehensive exploration of the next Nepali word prediction models showcasing the benefits of using LSTM and BiLSTM as well as the impact of contextual embeddings in enhancing predictive accuracy. This work shall contribute to advancing the field of NLP by improving the next word prediction models making strides towards more accurate and contextually aware language generation models in Nepali language and Nepali domain.

Keywords
BiLSTM, Contextual, LSTM, Nepali, Next Word, NLP

1. Introduction

Nepali, also known as Nepalese, is the official language of Nepal and is spoken by the majority of the population. Nepali belongs to the indo- Aryan branch of the Indo-European language family. It is closely related to languages like Hindi, Bengali and Punjabi. It is also recognized as one of the 22 scheduled languages of India and is spoken by significant communities in the Indian states of Sikkim, West Bengal and the Northeastern region[1].

All types of Language users, including new users, may write in any language with ease using a predictive system that accurately predicts the words. This type of system aids in teaching and learning any language that can prevent word errors by suggesting the correct words[2].

Predicting new words is dependent on the corpus and the type of corpus. It also depends on the size of the corpus. For example, a model trained on the dataset of sports may not predict the word of politics and the model trained on the dataset of politics may not predict the word of sports[3]. Also increasing the size of the corpus may lead to poor prediction as the model has to predict a few words from the large domain size and the predicting task depends on the probability of the words in the corpus. In this paper, it is focused on the prediction list and the accuracy of the prediction[4].

Natural Language Processing is the branch of AI that can understand verbal and written text in the same way as humans do. The RNN, LSTM, and BiLSTMs are used for their ability to remember long-range sequences. Recurrent neural networks of the Long Short-Term Memory (LSTM) type can recognise order relationships in sequence prediction problems[5].

A Bidirectional LSTM(BiLSTM) is a kind of RNN which is used for the task of NLP. It is an effective tool for modelling the sequential dependencies between words and phrases in both directions of the sequence since, in contrast to ordinary LSTM, the input flows in both directions and it may use information from both sides[6].

2. Literature Review

The use of Four-gram Language Model with the use of Viterbi algorithm has been tested and found to be more accurate than trigram and bigram[1].As the number of n-grams are increased, more accuracy is achieved as there are more repeated words.

N-gram language model used for next word suggestion system for the sorani and kurmanji dialects which is a part of kurdish language which is different from English language is an achievement for NLP[3].The Kurmanji and English language are written from left to right but sorani dialect is written from right to left. In the proposed system n-gram model was usedwhich was suitable for kurdish dialects. Whenever the word is not found in the dictionary or in case of lack of true evidence, the SBO algorithm was implemented which decreases the n-grams that could be used for next five word prediction.

The word prediction system's main goal is to save keystrokes, which is measured as the percentage of less keys pushed with a prediction than without any prediction at all[7].An analysis of the characteristics such as accuracy and failure rate for the unigram, bigram, trigram, backoff, and linear interpolation models is conducted. The results show that the linear interpolation model has the lowest failure rate while the trigram, backoff, and linear interpolation models have the highest accuracy. The researchers also discovered that the bigram model has the longest prediction length, which is five, while the other models have seven.

The next word prediction can be used for keystroke saving[8]. A keystroke's saving is a very challenging task. The use of next word prediction using trigram model can be easily used for 50% to 60% keystroke saving and exceeding this limit is quite difficult.

Next Word Prediction is a task of NLP fields as it is about mining the text[9]. The use of Deep Learning techniques like LSTM to make predictions has got an accuracy of 75% having the loss of 55% which is a good enough to predict next word.In this paper, a dictionary of hindi language has been used as training dataset and has trained the model using deep learning technique.

3. Methodology

3.1 RNN

A recurrent neural network (RNN) is a type of artificial neural network that is commonly used for sequence modeling[6]. RNNs can learn long-range dependencies in data, which makes them well-suited for tasks such as natural language processing, speech recognition, and machine translation. RNNs work by processing data one step at a time. At each step, the RNN takes in the current input and the previous hidden state, and it outputs a new hidden state. The hidden state is a representation of the RNN's understanding of the data up to that point.The RNN's hidden state is used to predict the next input. The prediction is made by a linear layer that is connected to the hidden state. The linear layer outputs a probability distribution over the possible next inputs.

3.2 LSTM

Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that is specifically designed to learn long-range dependencies in data. LSTMs are able to do this by using a gating mechanism that allows them to control the flow of information through the network[5]. The gating mechanism in an LSTM consists of three gates. The forget gate decides how much of the previous hidden state to forget. The input gate decides how much of the current input to add to the hidden state. The output gate decides how much of the hidden state to output. The forget gate, input gate, and output gate work together to control the flow of information through the LSTM. This allows the LSTM to learn long-range dependencies in data, even if the data is noisy or corrupted[8].

3.3 BiLSTM

A Bidirectional long short-term memory (BiLSTM) is a type of recurrent neural network (RNN) that is able to learn long-range dependencies in data from both the past and the future[7]. This makes it well-suited for tasks such as natural language processing, speech recognition and machine translation. BiLSTM works by processing data in two directions: from the past to the present and from the present to the future. This allows the BiLSTM to learn the context of the current word from both the previous words and the future words[8]. The BiLSTM is made up of two LSTMs: a forward LSTM and a backward LSTM. The forward LSTM processes the data from the past to the present and the backward LSTM processes the data from the present to the future. The outputs of the two LSTMs are then combined to form a single output.

3.4 Nepali Language

Nepali, also known as Nepalese, is the official language of Nepal and one of the constitutionally recognized languages of India. It belongs to the Indo-Aryan branch of the Indo-European language family[1]. With over 30 million native speakers, Nepali is widely spoken in Nepal and various Nepali-speaking communities across the world. Nepali is written in the Devanagari script, which is also used for many other South Asian languages, including Hindi, Sanskrit, and Marathi. The script consists of 36 consonant letters and a range of vowel diacritics that are combined to form syllables. Nepali is written from left to right.

3.5 Our Approach

3.5.1 Datasets

The dataset for the work is the primary dataset that is collected via different sources of information. The dataset is focused on the news data. Also, data were collected from social media posts, Nepali online newspaper articles on the field of politics, sports, economics and education. The summary of dataset is as follows:

Number of total words: 62474
Number of unique words: 15,099

3.5.2 Preprocessing

Once the data was collected, data cleaning process were carried out. For instance, all the punctuations symbols like (? , " % # @ & [ ] !) and all the Nepali digits were removed from the text and the file was saved in UTF-8 encodings. Then the tokenization process was carried out. Each individual word was assigned a unique number. Text to sequence operation

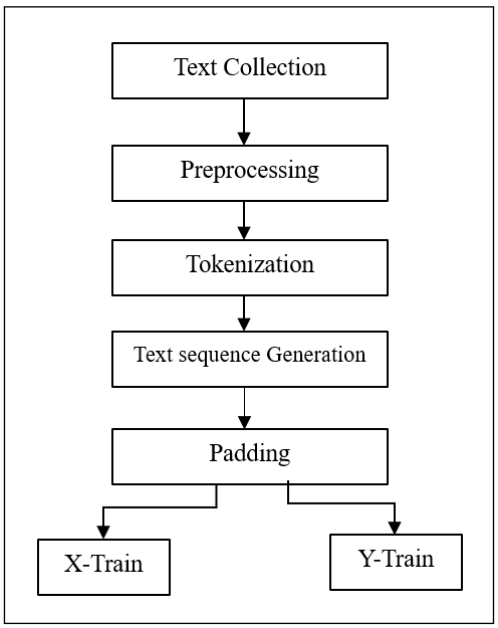

Figure 1: Text Preprocessing

was carried out to get the sequence of all the numbers of respective word that was present in the corpus. It generated an array of text sequences of length 62474. After text to sequence operation, an array of length 62471 was created which contained again arrays of numbers each of 4 sizes, and each array represented a sequence of number representing to their respective word. After this zero padding was applied and the array was converted to 2D array. To extract the features and label, first three elements were selected as the x-train and the fourth element of the pad sequence i.e. 2D array was selected as the y-train. The first three element was selected as x-train because the model is built on the basis that model expects first three word of the sentence and it recommends the fourth one. Thus, the x-train of size (62471,3) and y-train of size (62471,1) was achieved. Since it is a task of classification, the y-train was encoded using one hot encoding. After applying one hot encoding, the size of y-train formed was (62471,15100). The overall operations of data preprocessing can be visualized in Figure 1.

3.5.3 Architecture

The overall architecture of the model is shown in Figure 2

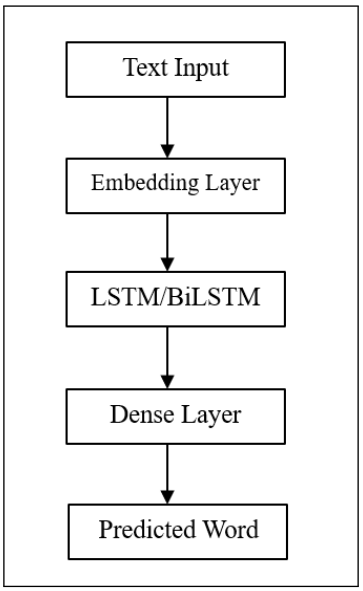

Figure 2: Architecture Of The Model

The input to the model is an array of numbers representing words. The first layer of the architecture is Embedding layer which takes number of vocabularies, vector size and length of each input. For the proposed model, number of vocabularies was 15100, and the vector size was 100. It means that each word of the input data was converted to a dense array of size 100. The input length was given as 3 as our x-train contains 3 numbers. The result of the embedding layer was fed as the input of LSTM and BiLSTM layers separately. Then a Dense layer with SoftMax as the activation function was used to predict the next word. The total number of units in the dense layer was made equal to the size of vocabulary. The output of the Dense layer was of size equal to number of unique words. From the resulted array, top 3 indices were selected that contained the highest probabilities values. The top three possible words is recommended by the model.

4. Results and Conclusion

After the successful training of the models with batch size of 16 and 7 epochs, a good accuracy for both LSTM and BiLSTM has been achieved and is summarized in the following table.

Table 1: Accuracy And Loss of LSTM And BiLSTM Model

	Accuracy	Val. Accuracy	Loss	val. Loss
LSTM	85.5	83.7	0.9334	1.1757
BiLSTM	92.33	91.43	0.3980	0.4374

The observations made while training the model shows that for LSTM model, the accuracy achieved was 85.5 percent and validation accuracy was found to be 83.70 percent. For BiLSTM, the accuracy achieved was 92.33 percent and the validation accuracy was found to be 91.43 percent.

The BiLSTM and LSTM Model has been tested for different text input and the result is shown below.

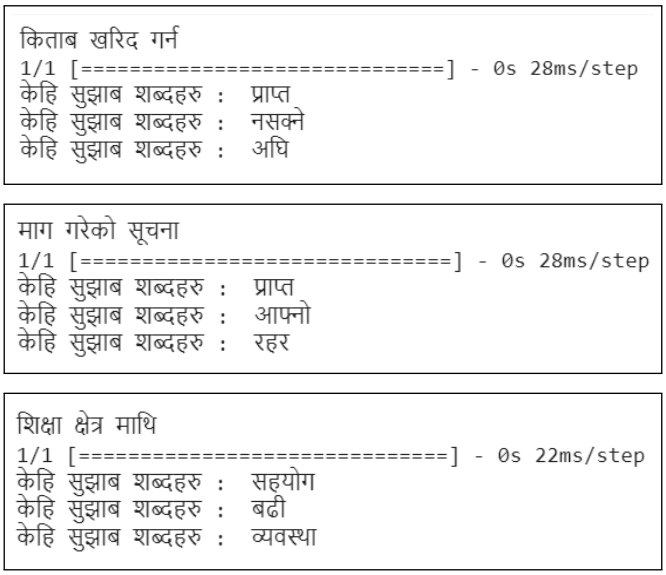

Figure 3: Text Prediction Using BiLSTM

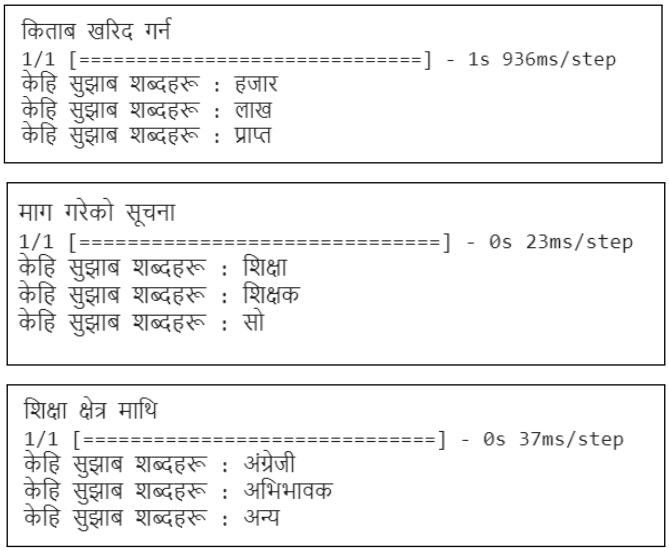

Figure 4: Text Prediction Using LSTM

The graph of model accuracy and model loss is shown in Figure 5 and Figure 6.

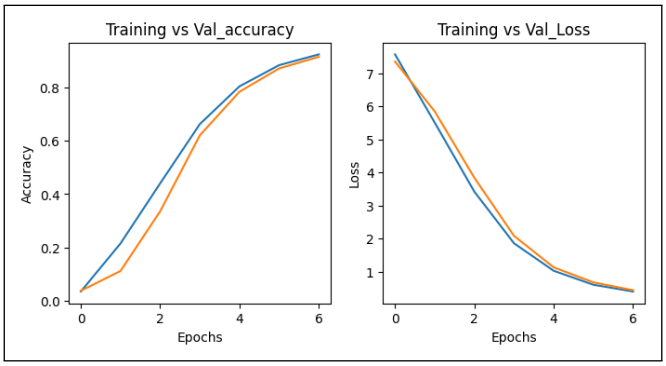

Figure 5: Graph of Accuracy and Loss of BiLSTM Model

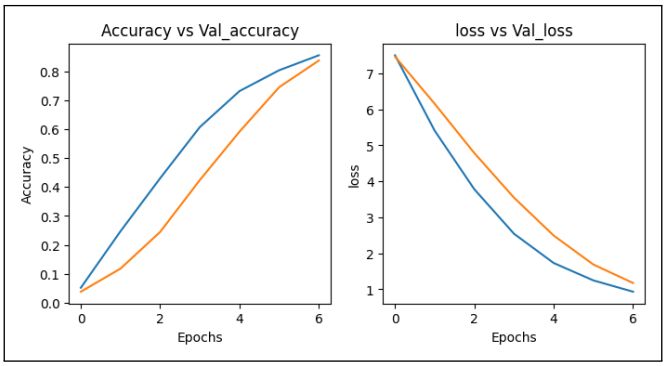

Figure 6: Graph of Accuracy and Loss of LSTM Model

The above graphs and findings demonstrate that the BiLSTM model outperforms the LSTM model in terms of accuracy and performance.

Acknowledgments

The authors would like to express their sincere gratitude to the Department of Electronics and Computer Engineering, IOE Purwanchal Campus Dharan for their collaboration in the initial stages of this research. Their expertise in the field greatly contributed to the formulation of the experimental design. The authors are very grateful to Asst. Proffessor Om Prakash Dhakal, Asst. Professor Sabin Kafley, Asst. Professor TN Jha for their insightful feedback and guidance throughout the research process, which significantly enhanced the quality of this work.We also extend our appreciation to all those who provided feedback during the peer-review process, contributing to the refinement of this paper.This work would not have been possible without the collective support and contributions of these individuals and organizations.

References

[1] Prithvi Narayan Campus. The use of n-gram language model in predicting nepali words. 2022.

[2] Felipe Zschornack Rodrigues Saraiva, Ticiana Linhares Coelho da Silva, and José Antônio Fernandes de Macêdo. Aspect term extraction using deep learning model with minimal feature engineering. In *Advanced Information Systems Engineering: 32nd International Conference, CAiSE 2020, Grenoble, France, June 8–12, 2020, Proceedings 32*, pages 185–198. Springer, 2020.

[3] Hozan K Hamarashid, Soran A Saeed, and Tarik A Rashid. Next word prediction based on the n-gram model for kurdish sorani and kurmanji. *Neural Computing and Applications*, 33:4547–4566, 2021.

[4] Masood Ghayoomi and Seyyed Mostafa Assi. Word prediction in a running text: A statistical language modeling for the persian language. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 57–63, 2005.

[5] Md Tarek Habib, Abdullah Al-Mamun, Md Sadekur Rahman, Shah Md Tanvir Siddiquee, and Farruk Ahmed. An exploratory approach to find a novel metric based optimum language model for automatic bangla word prediction. *International Journal of Intelligent Systems and Applications*, 12(2):47, 2018.

[6] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[7] Partha Pratim Barman and Abhijit Boruah. A rnn based approach for next word prediction in assamese phonetic transcription. *Procedia computer science*, 143:117–123, 2018.

[8] Orlando Iparraguirre-Villanueva, Victor Guevara-Ponce, Daniel Ruiz-Alvarado, Saul Beltozar-Clemente, Fernando Sierra-Liñan, Joselyn Zapata-Paulini, and Michael Cabanillas-Carbonell. Text prediction recurrent neural networks using long short-term memory-dropout. *Indones. J. Electr. Eng. Comput. Sci*, 29:1758–1768, 2023.

[9] Sanskriti Agarwal, Sukritin, Aditya Sharma, and Anurag Mishra. Next word prediction using hindi language. In *Ambient Communications and Computer Systems: Proceedings of RACCCS 2021*, pages 99–108. Springer, 2022.