Journal of Mines, Metals and Fuels, 71(5): 614-618; 2023. DOI: 10.18311/jmmf/2023/34159

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: www.informaticsjournals.com/index.php/jmmf

Assamese and BODO next Word Detection using LSTM

Ajit Das[1], Abhijit Baruah[2] and Sudipta Roy[3]

[1]Assistant Professor, Department of Computer Science and Technology, Bodoland University Kokrajhar, Assam 783370.
E-mail: adas0078@rediffmail.com
[2]Department of Computer Science and Technology Bodoland University Kokrajhar, Assam,783370.
E-mail: baruahabhijit1998@gmail.com
[3]Professor, Department of Computer Science and Engineering, Assam University Silchar, Assam 788011
E-mail: sudipta.roy@aus.ac.in

Abstract

The official language of the Indian state of Assam is Assamese, an Eastern Indo-Aryan language. Assamese, the sole native Indo-Aryan language in the Assam Valley, has been heavily impacted by the nearby Tibeto-Burman languages in terms of lexicon, phonetics, and grammar. Its grammar is renowned for its highly inflected forms, and both honorific and non-honorific formulations can include a variety of pronouns and plural nouns. Additionally closely linked to Bengali, Assamese lacks grammatical gender distinctions like Oriya and Bengali. On the other hand, The Bodo language is a variety of dialects of the Tibeto-Burman branch of the Sino-Tibetan languages. Assam, Meghalaya, and Bangladesh are all home to speakers of the Bodo language, which is spoken in northeastern India. It shares linguistic kinship with the Dimasa, Tripura, and Lalunga languages and is written in Bengali, Latin, and Devanagari scripts.

Another name for Next Word Prediction is Language Modeling. Predicting what word will be spoken right away requires commitment. It is one of the primary functions of NLP and has a wide range of uses. Our objective is to create this model as quickly and efficiently as possible. RNNs can interpret prior material and forecast words since they have a lengthy short-term memory, which might be useful for users when building sentences. This method creates words by using letter-by-letter prediction, or letter-by-letter prediction. Users can benefit from next word prediction, which makes typing faster and more accurate. The Assamese and Bodo languages rely on next word prediction since multiple characters can be created by pressing the same consonants combined with different vowels, vowel combinations, and special keys. As a result, we present a Long Short Term Memory (LSTM) network model for Assamese and Bodo next word prediction. With 63,300 point sentences, we test the suggested network model, and it achieves 96 per cent accuracy. In addition, we contrasted the suggested model with cutting-edge models like the LSTM. The proposed network model offers a promising outcome, according to experimental findings.

Keywords: *Assamese language, Bodo language, LSTM, NLP, Rnn*

1.0 Introduction

In the modern digital era, data input is one of the most crucial duties for those who use computers and other mobile electronic devices. Next word prediction during data entry speeds up text entry for people with low spelling abilities and enhances document generation by reducing spelling mistakes. For those with severe oral and motor difficulties, handwriting recognition, mobile phone use, or texting, next word prediction might be beneficial. Word sequence prediction may also be used as a data source in research on natural language processing, including those on voice and handwritten text. Ethiopia is increasingly depending on computers and mobile devices today. Both Bodo and

[*]Author for correspondence

Assamese are widely used for communication. The Semitic languages Bodo and Assamese are.

Natural language processing (NLP), which involves figuring out efficient ways to interact with and gain from human connections, is a significant part of artificial intelligence. One such dedication is to provide portable users with next word suggestions as they type in applications, with the ultimate goal of streamlining message transmission by letting the user choose a suggested word rather than compose it. Long-term short-term memory, or LSTM, allows for the comprehension of previously read text and word prediction. This method makes use of letter change prediction to determine whether a character will make up a word. This will help users write essays and organize long paragraphs faster and allow them to focus on the topic at hand rather than wasting time trying to figure out what to write next. We aim to design or simulate auto complete functions using LSTMs. The standard Nietzsche text file will serve as our training data for this LSTM experiment, which will test this problem. To carry out this task, the majority of software employs a variety of techniques, including standard neural networks and NLP. Next word prediction, also referred to as language modeling, is the process of anticipating the next word. It has a variety of applications and is a fundamental NLP task.

2.0 Related Work

The approach used to conduct the literature review is described in this section. This essay represents an examination of previous contributions to the academic community.

1. In this paper, the multi-window convolution (MRNN) algorithm is implemented. Additionally, residual-connected minimal gated unit (MGU), a condensed version of the long-short-term memory (LSTM), is created. By skipping a few layers during training, the cnn tries to reduce training time while maintaining high accuracy.
2. This paper used the RNN algorithm as well as GRU, another variant of RNN, for code completion because RNN aids in predicting additional code syntax. The authors contend that their approach is more geared toward current approaches. Next word prediction was divided into two parts: intra-vocabulary and identifier prediction. They predicted vocabulary words using an LSTM neural language model. The pointer network is identified using a pointer model.
3. The Bangla language was worked on by authors. For word completion and prediction, they suggested a novel approach. An N-Gram language model that predicts a set of words was their suggestion. They were successful in getting good outcomes.
4. The authors used LSTM to predict additional words for the Assamese language. They fed the transcribed language to their model and stored it in accordance with the International Phonetic Association (IPA) chart. They developed a prototype for those who are physically challenged. This model employs the Unigram, Bigram, Trigram method.
5. The author introduces a novel strategy for word prediction in the paper[5] that is based on contextual information and machine learning. This approach approached the issue as a learning classification task. The author trained the word predictor using a variety of feature selection methods for machine learning, including SVM (Support Vector Machine) with MI, X2, and others. Several data sets were used in the experiment, and when compared to other similar studies, this method's word prediction accuracy was found to be 91% accurate.
6. The authors of[6] proposed an LSTM model for the Assamese language that predicts the subsequent word(s) given a collection of recently used terms. There is no longer a problem with words having several synonyms because the Assamese language has been transliterated using the International Phonetic Association (IPA). Their model use LSTM to predict the following word with an accuracy of 88.20% while looking over a dataset of transliterated Assamese words. The result shows that LSTM is remarkably good at predicting the next word, even for a language with few resources.
7. An automatic code completion algorithm for the next word was presented in article[7]. Vocabulary word prediction and identifier prediction were the authors' two main strategies. A neural language model built on the LSTM network was proposed in order to predict words from the vocabulary. A model built on a network of indicators is suggested for the prediction of identifiers. The proposed method is assessed using the source code gathered in the online judging system. The experiment's findings demonstrate that the suggested method is highly accurate at predicting both the following word in the vocabulary and the following identifier.

3.0 Proposed Methodology/ Algorithm

An artificial neural network class used in deep learning and artificial intelligence is referred to as "LSTM." As contrast to conventional feed-forward neural networks, LSTM contains feedback connections. Such a recurrent neural network (RNN) can analyse single data points, like photos, as well as full data sequences, like audio or video. Examples of applications for LSTM include machine translation, connected, un-segmented handwriting recognition, speech

recognition, robot control, video games, and healthcare. In the 20th century, the LSTM neural network was the one that received the most references.

The term "long-term memory" (LSTM) is used to compare the "long-term memory" and "short-term memory" of a conventional RNN. Similar to how an instantaneous shift in electrical firing patterns in the brain maintains short-term memories, the activation patterns in the network change once each time step. Similar to how synaptic strength changes during physiological processes help the brain retain long-term memories, the weights and biases of connections in the network change once throughout each training event. The long-term short-term memory (LSTM), or long-term short-term memory, is something that RNNs are designed to have thanks to the LSTM architecture.

4.0 Methodology

The suggested next word prediction model is based on a set of Bodo sentences of varied lengths (number of tokens). To make short sentences the same length, we add a zero pad at the end. The encoded text is subsequently provided to the LSTM model's input layer. The suggested model's hyperparameters are established through experiments. Practice sentences help the model learn the proper word order. The model may be applied to programmes that create Assamese and Bodo words once it has been trained and stored. This saves time and reduces typing.

4.1 Dataset Overview and Data-Preprocessing

To assess the new strategy we have proposed, we have gathered a significant amount of data in Assamese and Bodo from TDIL (Technology Development for Indian Languages). The total amount of data collected is 63 thousand and has been gathered from various sources. After gathering the data from various sources, we cleaned it using the cleanup function, which was put in place to get rid of extraneous objects ("", (), /,!), and also helped to standardize the original dataset so that it could be used for other things in the future. Each time we want to predict a word or words, we typically input four words into our model. The output will always be one word no matter how many words are in the input word.

4.2 Tokenization

Tokenization is the process of breaking text into words. Tokenization can occur on any character, but the most common way to tokenize is to do it on a space character. Tokenization is the process of turning a string of letters into a string of tokens, which then has to be transformed into a string of numerical vectors that a neural network can understand.

4.3 System Flowchart

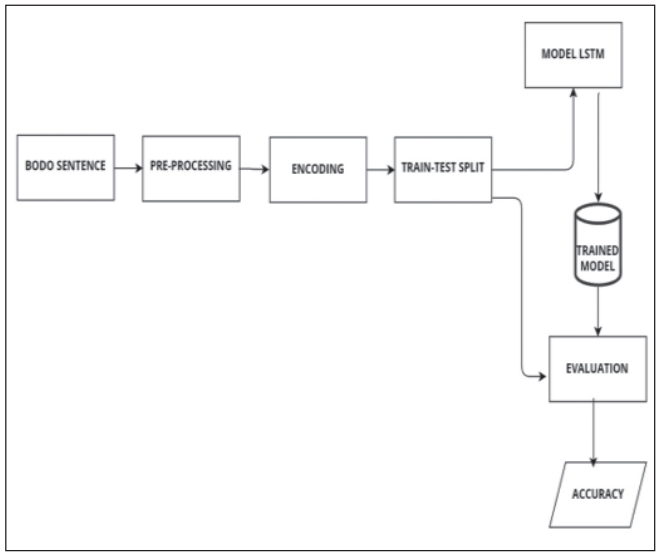


Figure 1

4.4 Model Architecture

Figure 3 depicts the structure of the trained model (Fig. No.1), which has two hidden layers named lstm and lstm_1 with a total of 17,621,753 (params). We used LSTM to train our previously created dataset and build one corresponding model.

One trained model for three length inputs is now available after training our dataset. The model uses three different word lengths as input and produces a single output, which is the next word that should most likely come after the input word sequence.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 3, 10)             45230

 lstm (LSTM)                 (None, 3, 1000)           4044000

 lstm_1 (LSTM)               (None, 1000)              8004000

 dense (Dense)               (None, 1000)              1001000

 dense_1 (Dense)             (None, 4523)              4527523

=================================================================
Total params: 17,621,753
Trainable params: 17,621,753
Non-trainable params: 0
```

Figure 2

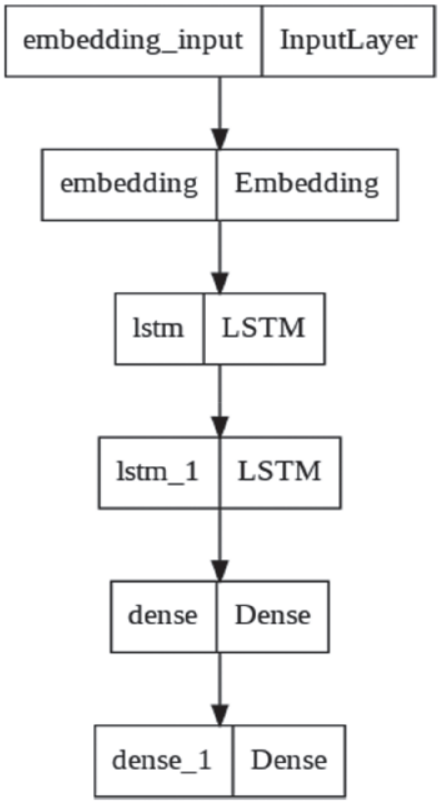


Figure 3

5.0 Result Analysis

Conducting experiments and carefully analyzing the results are required to validate the suggested approach. So, using a corpus dataset to train a model with identical structures for up to 70 epochs, we assessed our suggested method (Figure 2). The trained LSTM model has an average accuracy of 96 per cent, as shown in Figures 4 and 5.

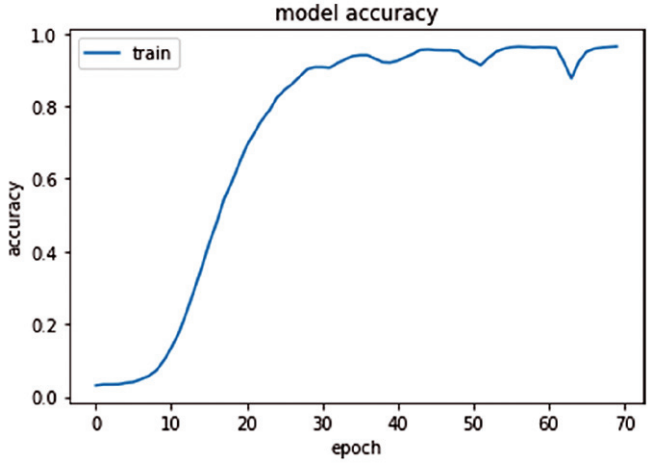


Figure 4

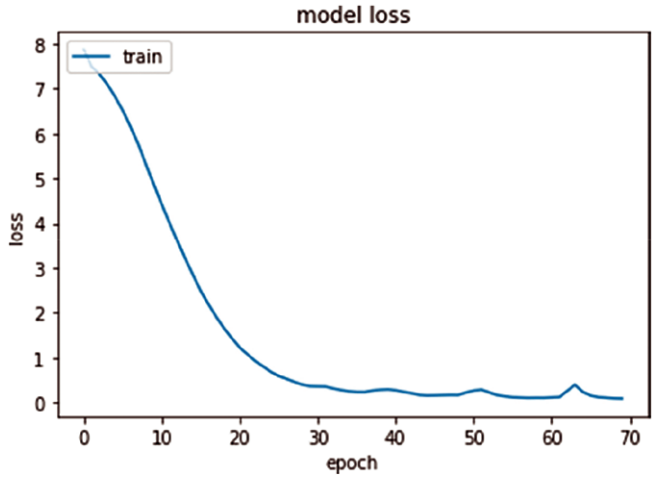


Figure 5

6.0 Conclusion

RNN has made a significant contribution to this research work by predicting the following best and most appropriate Assamese and Bodo word and sentence. If we could use the larger dataset we already used in this work, the overall accuracy of this approach would be even more impressive. It was difficult to use the Bodo corpus dataset because neither Assamese nor Bodo has a ready-made dataset, so we had to gather the data from various sources. To improve the RNN's ability to predict the next Assamese and Bodo word and sentence, we will work to amass a sizable dataset in the coming months. This study will also be useful as a tool for sustainable technologies in industry because of how widely it can be applied.

7.0 References

1. J. Yang, H. Wang and K. Guo, (2020): "Natural Language Word Prediction Model Based on Multi-Window Con-volution and Residual Network," in *IEEE Access*, vol. 8, pp. 188036-188043, doi: 10.1109/ACCESS.2020.3031200.
2. K. Terada and Y. Watanobe, (2019): "Code Completion for Programming Education based on Recurrent Neural Network," 2019 IEEE 11th International Workshop on Computational Intelligence and Applications (IWCIA), Hiroshima, Japan, pp.109-114, doi: 10.1109/IWCIA47330.2019.8955090.
3. Habib, Md AL-Mamun, Abdullah Rahman, Md Siddiquee, Shah Ahmed, Farruk. (2018): An Exploratory Approach to Find a Novel Metric Based Optimum Language Model for Automatic Bangla Word Prediction. *International Journal of Intelligent Systems and*

Appli-cations. 2. 47-54. 10.5815/ijisa.2018.02.05.

4. Partha Pratim Barman, Abhijit Boruah, (2018): A RNN based Approach for next word prediction in As-samese Phonetic Transcription, Procedia Computer Sci-ence, Volume 143, Pg. 117-123, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.10.359.
5. Al-Mubaid, Hisham. (2007): A Learning- Classification Based Approach for Word Prediction. *Int. Arab J. Inf.* Technol. 4. 264-271.
6. Pratim Barman Partha, Boruah Abhijit, (2018): "A RNN based Approach for next word prediction in Assamese Phonetic Transcription," Procedia Computer Science, vol. 43, pp. 117–123.
7. Terada Kenta and Watanobe Yutaka, (2021): "Code completion for programming education based on deep learning," *Int. J. Computational Intelligence Studies*, vol. 10, no. 2–3, pp. 109–114.
8. Sourabh Ambulgekar, Sanket Malewadikar, Raju Garande, and Dr. Bharti Joshi, (2021): Next Words Prediction Using Recurrent NeuralNetworks'', Ramarao Adik Institute of Technology Mumbai, Maharashtra, ITM Web of Conferences 40, 03034.
9. Omor Faruk Rakib, Shahinur Akter, Md Azim Khan, Amit Kumar Das, Khan Mohammad Habibullah, (2019): "Bangla Word Prediction and Sentence Completion Using GRU: An Extended Version of RNN on N-gram Language Model'', Department of Computer Science and Engineering East-West University Dhaka, Bangladesh, 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), 24-25 December, Dhaka.
10. Afika Rianti, Suprih Widodo, Atikah Dhani Ayuningtyas, Fadlan Bima Hermawan, (2022): "Next Word Prediction using LSTM'', *Journal of Information Technology and its Utilization*, Vol.5, issue 1, june, eissn 2654-802x.
11. Jordan, Michael I., and Tom M. Mitchell. (2015): "Machine learning: Trends, perspectives, and prospects." Science 349, no. 6245: 255-260.
12. Sahoo, Abhaya Kumar, Chittaranjan Pradhan, and Himansu Das. (2020): "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making." In Nature inspired computing for data science, pp. 201-212. Springer, Cham.
13. Prajapati, Gend Lal, and Rekha Saha (2019): "REEDS: Relevance and enhanced entropy based Dempster Shafer approach for next word prediction using language model." *Journal of Computational Science* 35 : 1-11.
14. Stremmel, Joel, and Arjun Singh. (2021): "Pretraining federated text models for next word prediction." In Future of Information and Communication Conference, pp. 477-18. 488. Springer, Cham.
15. Xiaoyun, Qu, Kang Xiaoning, Zhang Chao, Jiang Shuai, and Ma Xiuda (2016): "Short-term prediction of wind power based on deep long short-term memory." In 2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), pp. 1148-1152. IEEE, 2016.