ISSN 2976-1190 (Online)
Journal of Engineering and Sciences
Vol. 1 Issue 1

Nepali Spelling Checker

Bhawana Prasain[1], Nabin Lamichhane[2], Nabina Pandey[3], Prakriti Adhikari[4], Puja Mudbhari[5]

[1]bhawana.prs@gmail.com, [2]babunabin@wrc.edu.np, [3]nabina.pandey456@gmail.com, [4]adhikariprakriti3@gmail.com, [5]pujamudbhari1234@gmail.com

[1,2,3,4,5]*Department of Electronics and Computer Engineering, IOE, Pashchimanchal Campus, Tribhuvan University, Nepal*

(Manuscript Received 16/08/2022; Review: 15/09/2022; Revised 30/09/2022; Accepted 05/10/2022)

Abstract

Communication is a great deal in today's era. Be it composing emails, be it writing essays for higher education, be it maintaining relationships with your loved ones through text, be it searching for problems on the internet, we can't deny the fact that there's a chance of spelling mistakes. Misspelled words can provide wrong information to readers and may create a negative impact on them. So, a spell checker plays a significant role in this area. Spell Checker detects and suggests the misspelled words. Spelling errors run across all the languages, not only in English. Detecting and identifying misspelled words can be even more tedious in the context of the Nepali Language since not much research has been done in the past in comparison to the English Language. In this project, we have proposed a spell checker that detects and corrects the misspelled words in the Nepali Language, especially in the health domain. Nepali Spell Checker is a sequence-to-sequence model based on GRU which solves the vanishing gradient problem. GRU takes care of the information from the past without forgetting it through time and eliminates information that is irrelevant to the prediction. Our model is based on GRU which is trained for 10 epochs and gives an accuracy of 75.11%.

Keywords: Gated Recurrent Unit (GRU); Levenshtein Distance; Natural Language Processing (NLP); Nepali Spell Checker

1. Introduction

A spelling error makes the text harder to read, understand, and process. A writer might not have enough time or the ability to correct spelling errors manually. Automatic spelling correction (ASC) systems can be of great help. A spell checker can be either context-free or context-sensitive. A context-free spell correction system is a system where the wrong word does not depend on anything but itself. In this system, the wrong word does not care about the previous words, the word after it, or neither the total meaning of the sentence. This condition makes the system a little bit complex as less data is available to use for algorithms to predict the wrong word outcome. And in the case of a context-sensitive spell checker, the wrong word relies on its previous and next word in the sentence to learn the context of the sentence. Hence, we need a huge data set for training our model for better results.

There has not been enough research in the field of Nepali language spelling correction. In our project "contextual Nepali spelling correction" we build a system that is going to detect the spelling error in the Nepali text. Spell checker basically involves two steps. At first, the word that needs to be corrected is detected based on context. The context of the sentence to be corrected is determined by using the sequence-2-sequence model based on GRU neural network architecture. And secondly, the list of candidate words for correction is generated using the Levenshtein edit distance algorithm among which the best word is selected as the correct word for replacement with the erroneous word.

2. Methodology

Input dataset is preprocessed by removing punctuations and stemming is done in two steps. Tokens are created out of stemmed dataset. Each word is represented by a unique integer. From every sentence, using a windowing size of 2 and sequence length as 10, we create an input sequence and a label for each input sequence. Input sequence along with label is fed to GRU by using stratified K-fold to ensure homogenous distribution of dataset. Trainable embedding layer of output dimension 200 calculates relations between words. To prevent

overfitting, a dropout layer is added.20% of nodes are randomly dropped out in each pass and 80% of the nodes are used for training the model. 1024 layers are in used GRU. GRU provides a contextual model. Since target variables are integer encoded, we use sparse categorical cross entropy as loss function. Adam is used for optimization of neural network. SoftMax activation function in dense layer gives output words along with their corresponding probability of occurrence. Dense layer's output size is 1,20,000 which is equal to the vocab size of our dataset.

Levenshtein edit distance is used to select the best fit word in a given context for any word that does not exist in the vocabulary of the training set. GRU preserves the information from the past without forgetting it through time and eliminates information which is irrelevant to the prediction. It consists of two gates i.e. Reset Gate and Update gate.

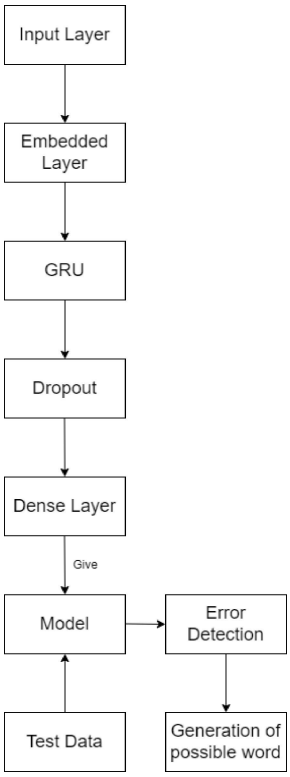

Figure 1 Flow chart for the model design.

3. Results and Discussion

For our Nepali spell checker model, we have used GRU and for measuring the performance of the sequence-to-sequence model we have used accuracy as metrics. Train set accuracy is 75.11% and test set accuracy is 73%.

4. Conclusions

This project addressed the problem of detection and correction of spelling errors in Nepali language. Our model works on two phases. During the first phase, the error word in the sentence is detected by using the seq2seq language model built using GRU deep neural network. After detection of the error word, in the second phase, Levenshtein distance algorithm is used for the generation of possible correct words also known as candidate words with edit distances 1, 2 and 3. Among the candidate words the words with the maximum probability of fitting in the sentence are given as suggestions to the user. And finally, the user can choose the best and correct form of word in their sentence. By performing training and testing of models in 10 epochs we obtained accuracy of 74%. By performing training and testing with more epochs, data and resources we can enhance the performance of our model.

Acknowledgment

We would like to express our profound gratitude to the National Conference on Recent Trends in Science, Technology and Innovation (RTSTI) for providing this platform to present our work. Besides, we would like to thank the Department of Electronics and Computer Engineering for the guidance and direction to accomplish this research-based project.

References

[1] Kukich K. Techniques for automatically correcting words in text, *ACM Computing Surveys*, 24(4) (1992) 377–439. doi: 10.1145/146370.146380.

[2] Arabic spelling correction using supervised learning. Online available: https://www.researchgate.net/publication/266319712_Arabic_Spelling_Correction_using_Supervised_Learning. [Accessed: 30-Nov-2021].

[3] Si-Lhoussain A., Gueddah H., and Yousfi A. Adaptating the Levenshtein Distance to Contextual Spelling Correction, *IJCSA,* 12(1) (2015) 127–133. url: http://www.tmrfindia. org/ijcsa/v12i111.pdf. [Accessed: 30-Nov-2021]

[4] Sharma S. and Gupta S. A correction model for real-word errors, *Procedia Computer Science,* 70 (2015) 99-106.
[5] Hu Y., Jing X., Ko Y., and. Rayz J. T. Misspelling Correction with Pre-trained Contextual Language Model, [2021]. arXiv:2101.03204 [cs], Jan. 2021, Accessed: Nov. 30. doi.org/10.48550/arXiv.2101.03204
[6] *Researchgate.net*. [Online]. Available: https://www.researchgate.net/publication/333300899_Real-Word_Errors_in_Arabic_Texts_A_Better_Algorithm_for_Detection_and_Correction. [Accessed: 2-Apr-2021].