A Thesis

on

«Nepali Text Part Of Speech Tagging Using Different Deep Learning Algorithms»

For Partial Fulfillment of the Requirements for the Degree of

Master of Computer Information System Awarded by

Pokhara University

Submitted by

Tika Ram Khojwar

MCIS

«Symbol No.: 202828»

«PU Reg. No.: 2020-2-92-0018»

Department of Graduate Studies

Nepal College of Information Technology

Balkumari, Lalitpur

«May, 2023»

1.3 Objectives of the Study

The best model depends on various factors, including the availability of training data, Language, computational resources, and the specific requirements of the application. So, experiment with different models and compare their performance on the specific task or dataset to determine the most suitable model.

The main objective of this paper is:

- Train and compare between different deep learning algorithms such as RNN, LSTM, BiLSTM and mBERT for nepali text POS tagging.
- Find out which algorithm is best suited for the process of POS tagging for Nepali text.

Chapter 2

LITERATURE REVIEW

There are only few researches have been done in the field of POS tagging for Nepali language. Some of them used statistical model (HMM) for identifying the tags while some used supervised machine learning model and some used supervised deep learning model to train the model.

Ashish Pradhan, Archit Yajnik (2021):

This article presents a comprehensive study and comparing two techniques, HMM and GRNN, for POS Tagging in Nepali text. The POS taggers aim to address the issue of ambiguity in Nepali text. Evaluation of the taggers is performed using corpora from TDIL (Technology Development for Indian Languages) which contains a total of 424716 tagged words with 39 tags, tags follow the guidelines of ILCI (Indian Languages Corpora Initiative), BSI (Bureau of Indian Standard), with implementation carried out using Python and Java programming languages, along with the NLTK Toolkit library. The achieved accuracy rates are as follows: 100% for known words (without ambiguity), 58.29% for ambiguous words (HMM), 60.45% for ambiguous words (GRNN), and 85.36% for non-ambiguous unknown words (GRNN). Although the GRNN tagger achieves the accuracy as high as the HMM Tagger, it fails or becomes unstable when the training data set is greater than 7000 words, while the HMM Tagger is trained with more than 400000 words with corresponding tags. A total of 4000 words are used for testing on both HMM and GRNN taggers.

Ingroj Shrestha, Shreeya Singh Dhakal (2021):

This article applied three deep learning models: BiLSTM, BiGRU, and BiLSTM-CRF for fine-grain POS tagging for the Nepali language. It uses Nepali National Corpus (NNC). It has 17 million manually and semi-manually words tagged with 112 POS-tags. Results show that deep learning models could capture fine-grained morphological features like gender, person, number, and honorifics that are encoded within words in highly inflectional languages like Nepali with a large enough dataset. This study trained all the models using two embedding: pre-trained multi-lingual BERT and randomly initialized Bare embedding. It found that training a randomly initialized Bare embedding is better than the ones trained

using large pre-trained multi-lingual BERT embedding for downstream tasks in Nepali like POS tagging. Among the tested models, the BiLSTM-CRF with the Bare embedding performed the best and achieved a new state-of-the-art F1 score of 98.51% for fine-grained Nepali POS tagging. This research contributes to the advancement of NLP techniques tailored specifically for the Nepali language.

Sarbin Sayami, Tej Bahadur Shahi and Subarna Shakya (2019):

This paper addresses the implementation and comparison of various deep learning-based POS taggers for Nepali text. The examined approaches include RNN, LSTM, GRU, and BiLSTM. These models are trained and evaluated using Nepali English parallel corpus annotated with 43 POS tag and contains nearly 88000 words which is collected from m Madan Puraskar Pustakalaya. The design of this Nepali POS Tag-set is inspired by the PENN Treebank POS Tag-set. The data set is divided into three sections i.e. training, development and testing. The accuracy obtained for simple RNN, LSTM, GRU and Bidirectional LSTM are 96.84%, 96.48%, 96.86% and 97.27% respectively. Therefore, Bi-directional LSTM outperformed all other three variants of RNN.

Archit Yajnik (2018):

This article focuses on POS tagging for Nepali text using the GRNN. Because GRNN is less expensive as compared to standard algorithms viz. Back propagation, Radial basis function, support vector machine etc. And also neural network is usually much faster to train than the traditional multilayer perceptron network. The corpus has total 7873 Nepali words with 41 tags. Out of which 5373 samples are used for training and the remaining 2500 samples for testing. The results show that 96.13% of words are correctly tagged on the training set, while 74.38% are accurately tagged on the testing set using GRNN. To compare the performance, the traditional Viterbi algorithm based on HMM is also evaluated. The Viterbi algorithm achieves classification accuracies of 97.2% and 40% on the training and testing datasets, respectively. The study concludes that the GRNN-based POS tagger demonstrates more consistency compared to the traditional Viterbi decoding technique. The GRNN approach yields a higher accuracy on the testing dataset, suggesting its potential for improved POS tagging in Nepali text compared to the Viterbi algorithm.

Archit Yajnik (2018):

The article that introduces POS tagging for Nepali text using three Artificial Neural Network (ANN) techniques. A novel algorithm is proposed, extracting features from the marginal probability of the Hidden Markov Model. These features are used as input vectors for Radial Basis Function (RBF) network, General Regression Neural Networks (GRNN), and Feed forward neural network. The training database contains 42100 words whereas the testing set consists of 6000 words with 41 tags. The GRNN-based POS tagging technique outperforms the others, achieving 100% accuracy for training and 98.32% accuracy for testing. This research contributes to Nepali POS tagging by presenting a novel algorithm and highlighting the effectiveness of the GRNN approach.

Archit Yajnik (2017):

This article focuses on POS tagging for Nepali text using the HMM and Viterbi algorithm. The study reveals that the Viterbi algorithm outperforms HMM in terms of computational efficiency and accuracy. Database is generated from NELRALEC Tagset with 41 tags. A report on Nepali Computational Grammar is made available by Prajwal Rupakheti et al. Database contains 45000 Nepali words with corresponding Tag, out of which 15005 samples are randomly collected for testing purpose. The Viterbi algorithm achieves an accuracy rate of 95.43%. The article also provides a detailed discussion of error analysis, specifically examining the instances where mismatches occurred during the POS tagging process.

Greeshma Prabha, Jyothsna P V, Shahina kk, Premjith B, Soman K P (2018):

This paper proposed a deep learning based POS tagger for Nepali text which is built using Recurrent Neural Network (RNN), Long Short Term Memory Networks (LSTM), Gated Recurrent Unit (GRU) and their bidirectional variants. It uses POS Tagged Nepali Corpus generated by translating 4325 English sentences from the PENN Treebank corpus tagged with 43 POS tags. The results demonstrate that the proposed model outperforms existing state-of-the-art POS taggers with an accuracy rate exceeding 99%. This research contributes to the field by showcasing the effectiveness of deep learning techniques in improving POS tagging for Nepali text.

Abhijit Paul, Bipul Syam Purkayastha, Sunita Sakar (2015):

This paper discusses HMM based POS tagging for the Nepali language. The study evaluates the HMM tagger using corpora from Technology Development for Indian Languages (TDIL) which contains around 1,50,839 tagged words and tagset consists of 42 tags including generic attributes and language specific attribute values. It has been followed the guidelines of ILCI (Indian Languages Corpora Initiative), BSI (Bureau of Indian Standard). The implementation is done using Python and the NLTK library. The HMM-based tagger achieves an accuracy of over 96% for known words but the system is not performing well for the text with unknown words yet. Overall, the paper provides insights into the effectiveness of HMM for Nepali POS tagging and highlights areas for future improvement.

Tej Bahadur Shahi, Tank Nath Dhamala, Bikash Balami (2013):

This paper focuses on SVM based POS tagger for Nepali language which uses the dictionary as a primary resources. This dictionary is collected from the FinalNepaliCorpus which contains only 11147 unique words. The POS tagging approaches like rule-based and HMM cannot handle many features that would generally be required for modeling a morphologically rich language like Nepali. SVM is efficient, portable, scalable and trainable. So, this paper proposes a SVM based tagger. The SVM tagger constructs feature vectors for each word in the input and classifies them into one of two classes using a One Vs Rest approach. The SVM algorithm achieves an accuracy rate of 96.48% for known words, 90.06% for unknown words and 93.27% in overall. That means SVM tagger demonstrates strong performance for known words. In comparison to rule-based and Hidden Markov Model (HMM) approaches, the SVM-based tagger exhibits a slightly higher overall accuracy.

Chapter 3

RESEARCH METHODOLOGY

In this chapter, we present the proposed method for determining the POS tags of the provided text. The following figure provides an overview of the tagging process.

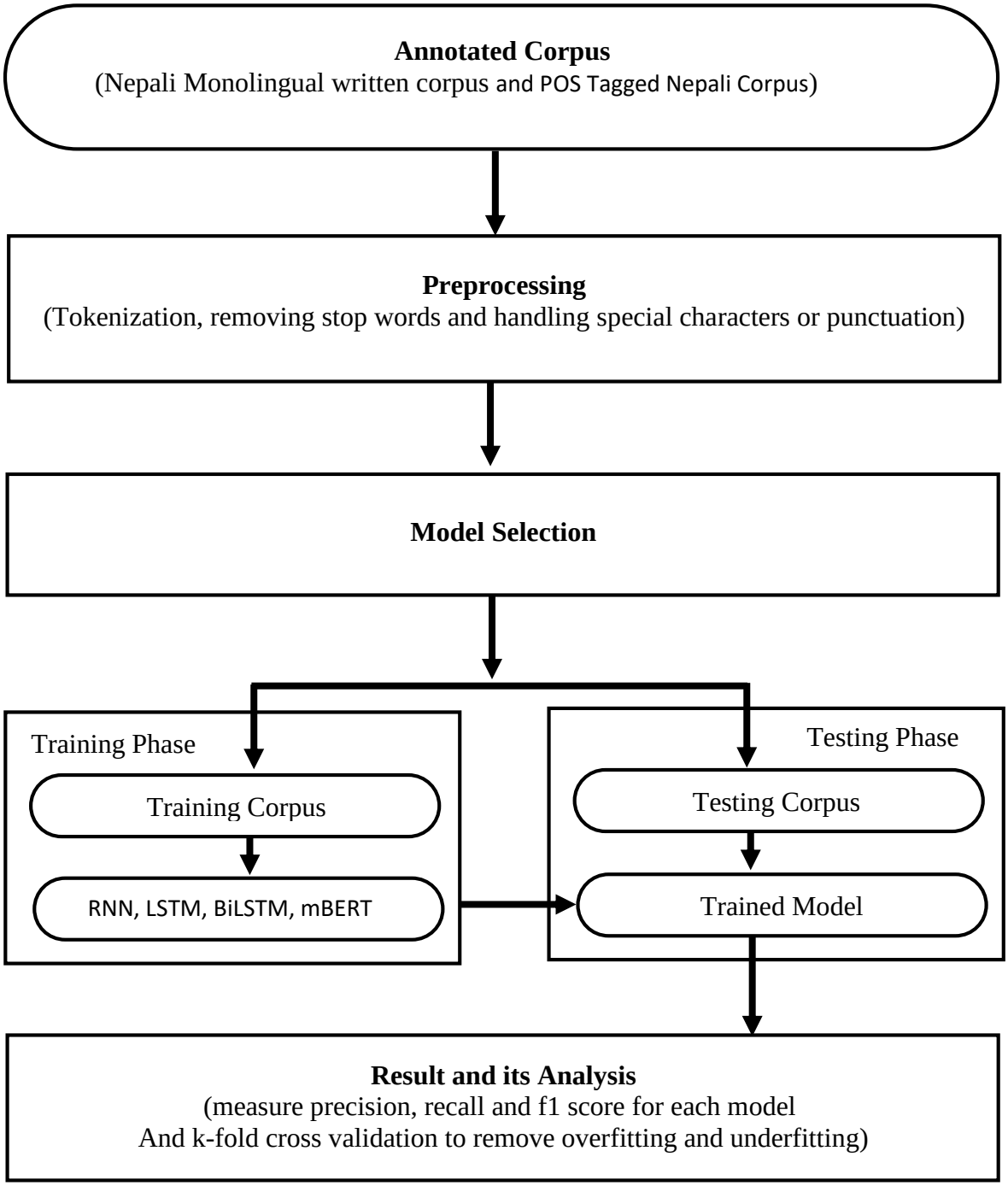

Fig 3.1: Proposed methodology for POS tagging

3.1 Dataset Collection

This research study will implement two corpus one is Nepali Monolingual written corpus. It consists of two main parts: the core corpus (core sample) and the general corpus. The core sample (CS) is a compilation of Nepali written texts from 15 diverse genres, with each text containing 2000 words. These texts were published between 1990 and 1992. On the other hand, the general corpus (GC) comprises written texts gathered from various sources, including the internet, newspapers, books, publishers, and authors, opportunistically collected without a specific sampling criteria. The corpus has total 2,202,000 words. It is a morphologically annotated corpus. A parts-of-speech tagset has been produced within the project: the Nelralec Tagset.

The other is POS Tagged Nepali Corpus which is available in Center for Language Engineering (CLE) website. It contains 4325 Sentences with 100720 annotated words. The POS tagset has 43 POS tags (null included) which is influenced by the PENN Treebank tagset.

3.2 Preprocessing Dataset

Only good dataset can give good output. To make good dataset, we need to transform the text into something meaningful that the algorithm can use. Preprocessing includes the tokenization, removing less useful parts such as removing stop words, empty line etc, stemming or lemmatization, and handling special characters or punctuation. Text data are converted to label encoded form to represent numerically so that models can perform well.

3.3 Model Selection

There are several models have been widely used and achieved good performance in POS tagging tasks. Different models have their own different features and specific task. There is no single "best" model for POS tagging, as the effectiveness of a model can vary depending on factors such as the dataset, language, and specific requirements of the task. According to previous research Deep Learning algorithms based models gives better accuracy in testing dataset and can deal with ambiguous, unknown words as compare to rule based, statistical and machine learning algorithms for POS tagging. Let’s see some probable DL model for labeled dataset.

3.3.1 RNN

First selected model is Recurrent Neural Network (RNN). It is a type of neural network specifically designed to handle sequential data. It has feedback connections that enable it to maintain and utilize information from previous steps in the sequence. RNNs are effective in capturing dependencies over time and are commonly used in tasks such as natural language processing, speech recognition, and time series analysis. They can learn patterns and make predictions based on the context of the sequence. Overall, RNNs are powerful tools for modeling and processing sequential data.

3.3.2 LSTM

Second selected model is Long Short-Term Memory (LSTM) which is a type of RNN architecture. It can deal with capturing long-term dependencies in sequential data. Because LSTMs utilize a memory cell, along with input, forget, and output gates, to selectively retain or discard information based on context. This helps them overcome the vanishing gradient problem and effectively learn from sequences of varying lengths. LSTMs have been successfully applied to tasks such as natural language processing, speech recognition, machine translation, and time series analysis. They are known for their ability to capture long-term dependencies and have become popular for modeling sequential data.

3.3.3 BiLSTM

Third selected model is Bidirectional Long Short-Term Memory (BiLSTM) is a variant of the LSTM model that processes input sequences in both forward and backward directions. By capturing context from both past and future elements, Bi-LSTM can model long-range dependencies and excel in tasks that require complete context understanding. It is widely used in NLP tasks, enabling improved performance and accuracy in analyzing sequential data with bidirectional context.

3.3.4 mBERT

Multilingual Bidirectional Encoder Representations from Transformers (mBERT) is a variation of the BERT model that is trained on multiple languages simultaneously. It is designed to handle multilingual and cross-lingual NLP tasks. mBERT is trained on a large corpus of text data from different languages, allowing it to learn shared representations that capture similarities and transferable knowledge across languages. It learns deep contextual representations of words and can effectively handle tasks such as text classification, part-

of-speech tagging, named entity recognition, and machine translation across multiple languages.

One of the key advantages of mBERT is that it can perform well on a wide range of languages without requiring separate models for each language. It can leverage shared representations to transfer knowledge from high-resource languages to low-resource languages, improving performance on languages with limited training data.

3.4 POS Tagging

After model selection training phase will start and then train the selected models RNN, LSTM, BiLSTM, and BERT by using Training dataset that is split form the original dataset. After the models are trained, they will test with the test dataset in the testing phase. The trained POS tagger will fed with the sentences as input and the corresponding POS tags as the expected outputs.

3.5 Results and its analysis

In this step, we will check whether the predicted tag and the expected tags are same. We will make a count on how many of the words will be correctly tagged and how many will be falsely tagged. Based on these we will calculate the accuracy of our model. The precision, recall and f1 score will also be measured for each model. To remove the issue of overfitting and underfitting, cross validation technique will be also used, where the dataset will divide in the three folds and in each iteration two of the folds will be taken as training set and the remaining one will be taken as the testing set.

These parameters are used to compare the performance of the implemented models.