Neural POS Tagger

NLP PROJECT:

TIRTH PANDIT (2019201017)

PRATIK TIWARI (2019201023)

MENTOR: UJWAL NARAYAN

GITHUB REPO: HTTPS://GITHUB.COM/PRATIKIIITH/NEURAL-POS-TAGGER-NLP.GIT

Problem Statement

- POS tagging is the process of tagging the words with their categories that best suits the definition of the word as well as the context of the sentence. It is often the first step for many NLP applications.
- Aim of the project is to create the Neural network based POS Tagger for Gujrati Language.
- In general and Gujrati Language in particular is not a very widely explored language in NLP Tasks. Also morphological complexity of the language makes it hard to develop NLP applications around it.
- Previous Work has been done in the area of POS Tagging in Gujrati Language which uses different approaches like HMM (Hidden Markov Models) and CRF(Conditional Random Fields)
- Purpose of this Project also includes the comparisons between these classical approaches and Neural POS Tagger

Dataset

- Gujarati Monolingual Text Corpus ILCI-II
- This is the final outcome of the project and there are approx. 30,000 sentences of general domain.
- The translated sentences have been POS tagged according to BIS (Bureau of Indian Standards) tag set.
- https://www.ldcil.org/Download/Tagset/LDCIL/5Gujrati.pdf
- It has eleven primary tags and similarly it divides in sub tags. Main classes of Tag-set are:
  - ✓ Demonstrative (DM) , Post Position (PSP) , Noun (NN)
  - ✓ Adverb (RB) , Verb (V) , Pronoun (PR)
  - ✓ Conjunction (CC) , Particles (RP) , Quantifier (QT)
  - ✓ Adjective (JJ) , Residual(RD).

Baseline Model

In this phase our aim is to implement and test the HMM and CRF Model for pos tagging the Given Dataset . Following research papers are used as the reference.

1. Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields (Chirag Patel and Karthik Gali)
   This paper describes a machine learning algorithm for Gujarati Part of Speech Tagging. The machine learning part is performed using a CRF model.

2. A Statistical Method for Evaluating Performance of Part of Speech Tagger for Gujarati (Pooja M. Bhatt, Amit Ganatra)

   This article offers POS tagging for Gujarati textual content the use of Hidden Markov Model. Using Gujarati text annotated corpus for training checking out statistics set are randomly separated. 80% accuracy is given by model.

3. Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge (Manish Shrivastava ,Pushpak Bhattacharyya)
   This paper presents a simple HMM based POS tagger, which employs a naive(longest suffix matching) stemmer as a pre-processor to achieve reasonably good accuracy of 93.12%. This method does not require any linguistic resource apart from a list of possible suffixes for the language. This list can be easily created using existing machine learning techniques.

1. POS Tagging For Resource Poor Indian Languages Through Feature Projection (Pruthwik Mishra, Vandan Mujadia, Dipti Misra Sharma)
   This Paper describes POS tagging without any labeled data. Our method requires translated sentences from a pair of languages. We used feature transfer from a resource rich language to resource poor languages.

Neural based Model

In this phase our aim is to Implement Neural Network Based POS Tagger , and compare results of Previous models and Neural based Model. Following Research Papers are used for the Reference

1. Character-level Supervision for Low-resource POS Tagging

   Paper presents an architecture for learning more robust neural POS taggers by jointly training a hierarchical, recurrent model and a recurrent character based sequence-to-sequence network

2. A Neural Network Model for Part-Of-Speech Tagging of Social Media Texts

   This paper presents a neural network model for Part-Of-Speech (POS) tagging of User-Generated Content (UGC) such as Twitter, Facebook and Web forums. The proposed model is end-to-end and uses both character and word level representations. Character level representations are learned during the training of the model through a Convolutional Neural Network (CNN). For word level representations, we combine several pre-trained embeddings (Word2Vec, FastText and GloVe).

3. Part-Of-Speech Tagging using Neural network by Ankur Parikh

   This paper presents two novel approaches of POS tagging using Neural network for Hindi language and compares them with two other machine learning approaches, HMM and CRF. In this paper, a single-neuro tagger, a Neural network based POS tagger with fixed length of context chosen empirically is presented first. Then, a multineuro tagger which consists of multiple single-neuro taggers with fixed but different lengths of contexts is presented

Neural based Model

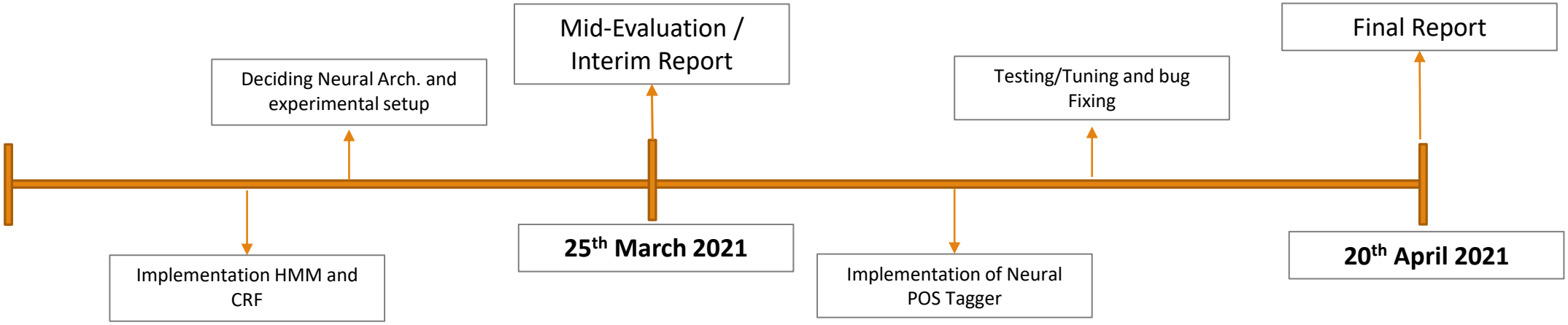
4. Neural Network based Parts of Speech Tagger for Hindi

In this paper, Artificial Neural Network for Hindi parts of speech tagger has been used. Uses Rule base POS Tagger as the initial classifier and on top of it trains the Neural network to finally classify the words in tag classes.

5. GENERAL REGRESSION NEURAL NETWORK BASED POS TAGGING FOR NEPALI TEXT

This article presents Part of Speech tagging for Nepali text using General Regression Neural Network (GRNN). The corpus is divided into two parts viz. training and testing. The network is trained and validated on both training and testing data. It is observed that 96.13% words are correctly being tagged on training set whereas 74.38% words are tagged correctly on testing data set using GRNN

Project Timeline



Final Deliverables

1. HMM and CRF implementation + result metrics on given dataset
2. Classical approaches + Neural POS tagger implementation (LSTM or GRU or Transformers)
3. Comparison Metrics for baseline and Neural based model.
4. Project Report