POS Tagger

Project Outline

Group Members:
Dharmeshgiri 2019201025
Smitkumar 2019201021

Project Mentor:
Ujwal Narayan

Project Definition

- Create a Statistical or Neural Part of Speech(POS) Tagger.
- It is the process of assigning correct part of speech to each word of a given input text depending on the context.
- Given table contains example of different tag and it's meaning in part of speech tagging.

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

DATASET : Gujarati

- This dataset is prepared by JNU Delhi in Gujarati language.
- It has 30,000 sentences.
- Link : https://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1882&lang=eni

Baseline

- **HMM** (**Hidden Markov Model**) is a Stochastic technique for **POS tagging**.
- Hidden Markov models are known for their applications to reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, musical score following, partial discharges, and bioinformatics.
- 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.

Neural Part

- Neural Based Models for PoS Tagging.
  - RNN
  - LSTM
  - Bidirectional LSTM

Feature Selection

- The prefix characters up to 7 characters
- The suffix characters up to 4 characters
- Length of the word
- Context Window size of 3 (Previous word, Current word and Next word)

Research Paper

1. Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge
   a. https://www.cse.iitb.ac.in/~pb/papers/icon08-hindi-pos-tagger
   b. In this paper, They present 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.

Research Paper

2. POS Tagging For Resource Poor Indian Languages Through Feature Projection
    a. https://www.researchgate.net/publication/323174678_POS_Tagging_For_Resource_Poor_Indian_Languages_Through_Feature_Projection
    b. They used feature transfer from a resource rich language to resource poor languages. Across 8 different Indian Languages, they achieved encouraging accuracies without any knowledge of the target language and any human annotation For Indian Languages, they considered the following morph features The prefix characters up to 7 characters, The suffix characters up to 4 characters, Length of the word, Context Window size of 3 (Previous word,Current word and Next word)

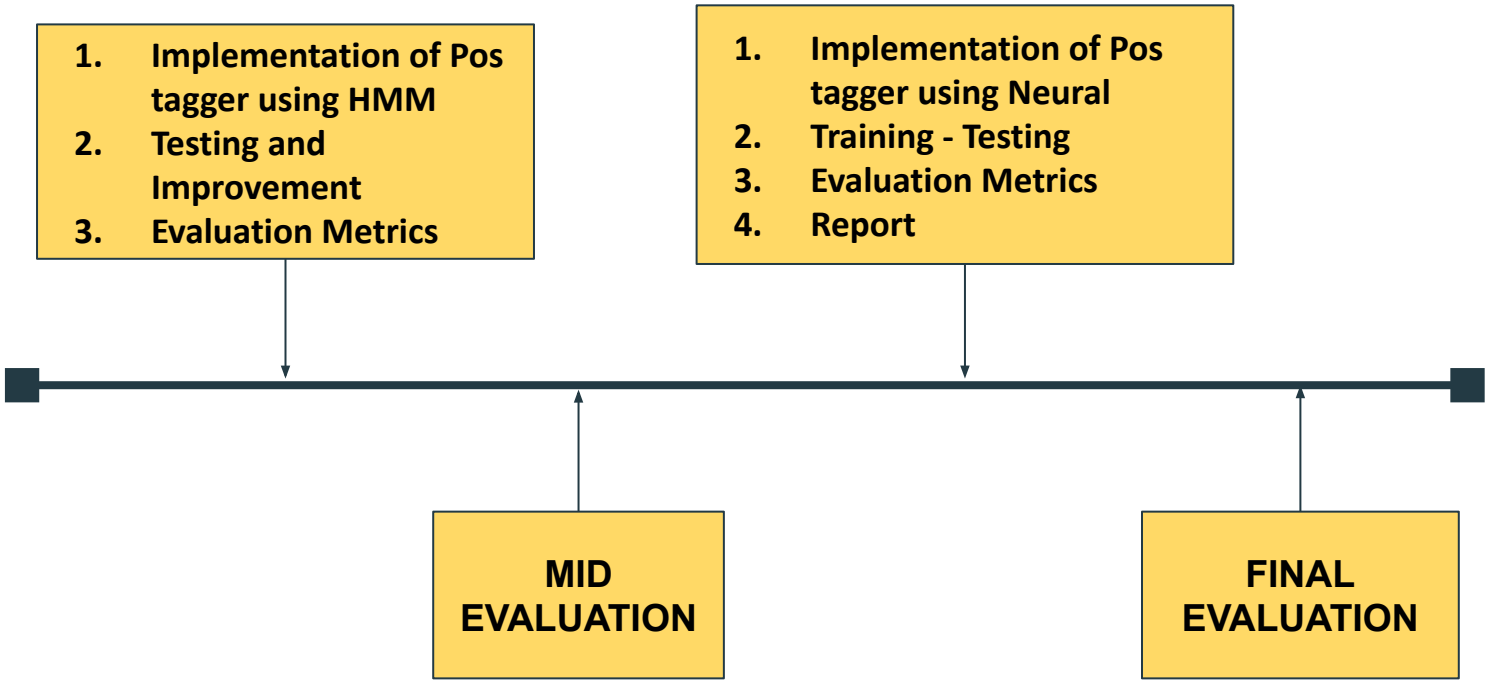
Research Paper

3. Character-level Supervision for Low-resource POS Tagging
    a. https://pdfs.semanticscholar.org/a93a/3799ba977aa2393cad8cd260d6f778a495e0.pdf?_ga=2.249332687.936938462.1614443610-381175129.1614443610
    b. In this paper experiment with three auxiliary tasks: lemmatization, character-based word autoencoding, and character-based random string autoencoding. They have used bidirectional LSTM.

Research Paper

4. Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields
    a. https://www.aclweb.org/anthology/I08-3019.pdf
    b. They used dataset where where the training corpus is of 10,000 words and the test corpus is of 5,000 words in Gujarati. The algorithm has achieved an accuracy of 92%. . The machine learning part is performed using a CRF model.

Project Timeline



Final Deliverables

- Implementation of HMM as baseline for POS Tagging.
- Implementation of Neural Model for POS Tagging.
- Evaluation Metrics for both models.
- Comparison analysis between baseline and neural model.
- Final project report.

THANK YOU