Summary of Part-of-Speech Tagging on Regional Language

Madhuri M. Deshpande et al. [1] , In this work, multi pass hybrid Part-Of-Speech(POS) tagger is built for the Marathi language sentences. A feature vector is build for every word in the sentence by referring to the next and the previous word i.e succeeding and preceding word that is tagged. The input sentence in Marathi is analyzed through the tokenization of every word in sentence and later stem for the every token is found. Each token is then analyzed for its tense, aspect, mood and POS tag. This approach is known as POS tagging .There may be a chances of ambiguities in the tagging process. This Marathi POS tagger uses rule-base, lookup files and dictionary. For every input sentence rule-base will provide a rules. The tagger utilizes lookup files and dictionary to list suffixes. After identifying the suffixes, the stem are obtained through removal of the suffixes and the applying the morpheme sequence rule. To obtain the correct POS tags and stem , dictionary has to be exhaustive and updated time to time. The accuracy achieved with this hybrid POS tagger is 84%.

Suraksha N M et al. [2], Part-Of-Speech tagging is considered to be the second step in Regional Natural language Processing. In this work POS tagging and Chunking is done using Conditional Random Fields(CRFs). The kannada corpus is used of 3000 sentences. The allocation of the corpus is done with respect to training set and the test set. For Training 2500 sentences are used and for testing 500 sentences. CRFs is the Probabilistic approach for applying sequence labeling task for the POS tagging for a Kannada text. Conditional random fields can be utilize in segmenting and labeling data sequentially. CRFs applied in different applications as well like named entity reognition, shallow parsing , chunking etc. The accuracy achieved with this model is 96.86% .

Sanjeev Kumar Sharma et al.[3], Part-of-Speech tagging is the technique of assigning tag to every word in the sentence. In this paper work Bi-gram Hidden Markov Model is used to improve accuracy of the existing Punjabi POS tagger which is rule based approach. For the Estimation of HMM parameters and training of the model a corpus of 20,000 words are used. The approach used in the estimation of the HMM parameters is Maximum likelihood. Viterby algorithm is used in the implementation for the HMM approach. The accuracy achieved using this model is 90.11%

References:

[1] Madhuri M. Deshpande, Dr. Sharad D. Gore ,” A Hybrid Part-of-Speech Tagger for Marathi Sentences”, Department of Computer Science, Savitribai Phule Pune University, 2018 International Conference on Communication, Information & Computing Technology (ICCICT), Feb. 2-3, Mumbai, India

[2] Suraksha N M, Reshma K , Shiva Kumar K M , “Part-Of-Speech Tagging And Parsing Of Kannada Text Using Conditional Random Fields (CRFs)” Department of Computer Science, Amrita University Mysuru, India, 2017 International Conference on Intelligent Computing and Control (I2C2)

[3] Sanjeev Kumar Sharma, Dr. Gurpreet Singh Lehal, “Using HIDDEN MARKOV MODEL to Improve the Accuracy of Punjabi POS Tagger” BIS college of Engg and Technology Moga, India, Department of computer science Punjabi University Patiala, India