Volume 7, Issue 4, April 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Parts of Speech Tagging for Indian Languages Review and Scope for Punjabi Language

[1]Ramandeep Kaur[*], [2]Lakhvir Singh Garcha, [3]Dr. Mohita Garag, [4]Satinderpal Singh

[1, 2, 3] North West Institute of Engineering &Technology, Moga, India

[4] Sri Guru Granth Sahib World University, Fatehgarh Sahib, India

Abstract— Part-of- Speech tagging is the way to tag every word in a text as a particular part of speech, e.g. proper verb, adverb etc. POS tagging is the first important step in the processing of NLP applications. This paper reports the survey on POS tagging for various Languages. Various techniques used for POS tagging also described in this paper. Due to complex structural effect, the number of problems occurs when tagging the sentences written in various languages. A lot of work has been done by the researchers in this field for various languages using various techniques HMM (Hidden Marcov Model), SVM (Support Vector Machine), ME (Maximum Entropy) etc.

Keywords— Natural Language Processing, Part of speech Processing, Tagset, Indian Languages

I. INTRODUCTION

The NLP (natural language processing) is the process that provides the facility of interaction between human and machine. It is a component of computer science, linguistics and artificial intelligence. It is difficult task to build NLP application because human speech is not always specific. The main objective of NLP is to develop such a system that can understand text and translate between human language and another. The work in area of Part-of-Speech (POS) tagging has begun in the early 1960s. Part of Speech tagging is an important tool for NLP. It is one of the simplest as well as statistical models for many NLP applications. POS Tagging is an initial step of information extraction, summarization, retrieval, machine translation, speech conversion [2].

POS tagging is the process of assigning the best grammar tag to each word of text like verb, noun, pronoun , adjective , adverb, conjunction , preposition etc. some unknown words exist in every language so it is very difficult task to assign the appropriate POS tag to each word in a sentence [3]. The mostly work that has been done for Indian languages was one of the rule based approaches and other empirical based POS tagging Approach. But the fact was that rule-based approach requires proper language knowledge and hand written rule. Due to morphological effect of Indian languages, researchers faced a great problem to write proper linguistic rules and many cases it was noticed that results were not good. Most of natural language processing work has been done for Hindi, Tamil, Malayalam and Marathi and several part-of-speech taggers have been applied for these languages. After this, researchers moved to stochastic based approach. However the stochastic methods requires large corpora to be effective, but still many successful POS were developed and used in various natural language processing tasks for Indian language. The main issue after morphological richness of Indian Languages is Ambiguity. It is very time consuming process to assign a correct POS tag to different context words. Due to this reason, POS Tagging is becoming a challenging problem for study in the field of NLP [1].

II. LITERATURE SURVEY FOR INDIAN LANGUAGES

In this In this paper (Kumar D., 2010) Antony P J and Dr. Soman had presented a survey on developments of different POS tagger systems as well as POS tagsets for Indian languages and the existing approaches that have been used to develop POS tagger tools . They concluded that almost all existing Indian language POS tagging systems are based on statistical and hybrid approach. This Paper (Antony P. J, 2011) specifies A CRF (Conditional Random Fields) based part of speech tagger and chunker for Hindi had been used by Aggarwal Himashu and Amni Anirudh. After evaluation they found that the strength of Conditional Random Fields can be seen on large training data and CRF performs better for chunking rather than for POS tagging with the training on same sized data. With training on 21000 words with the best feature set, the CRF based POS tagger is 82.67% accurate, while the chunker performs at 90.89% when evaluated with evaluation script from conll 2000.

In this paper (Kaur M, 2014) Navneet Garg, Vishal Goyal, Suman Preet used Rule Based Hindi Part of Speech Tagger for Hindi. The System is evaluated over a corpus of 26,149 words with 30 different standard part of speech tags for Hindi. The evaluation of the system is done on the different domains of Hindi Corpus. These domains include news, essay, and short storie and system achieved the accuracy of 87.55%.

This paper (M & UzZaman N, 2007) specfies A Comparison of Unigram, Bigram, HMM and BrillâA§s POS Tagging ˘ Approaches for some South Asian Languages has been done by Fahim Muhammad Hasan compared the performance of n-grams, HMM or transformation based POS Taggers on three South Asian Languages, Bangla, Hindi and Telegu. And we found that the HMM based tagger might perform better for English, but for South Asian languages, using corpora of different sizes, the transformation based BrillâA§s ˘ approach performs significantly better than any other approach when using a 26-tags tagset and pre-annotated training corpora consisting of a maximum of 25426, 26148 and 27511 tokens for Bangla, Hindi and Telegu respectively.

In this paper (Mohnot K & Singh S P, 2014) Manjit Kaur , Mehak Aggerwal and Sanjeev Kumar Sharma introduced an improving Punjabi Part of Speech Tagger by Using Reduced Tag Set. They Effort to improve the accuracy of HMM based Punjabi POS tagger has been done by reducing the tagset. The tagset has been reduced from more than 630 tags to 36 tags. We observed a significant improvement in the accuracy of tagging. Their proposed tagger shows an accuracy of 92- 95% whereas the existing HMM based POS tagger was reported to give an accuracy of 85-87%.

In this paper (Mahar J A, 2010) Nisheeth Joshi1, Hemant Darbari and Iti Mathur described efforts to build a Hidden Markov Model based Part of Speech Tagger. They used IL POS tag set for the development of tagger. HMM based statistical technique was used to train POS tagger for Hindi. They disambiguated correct word-tag combinations using the contextual information was available in the text and attained the accuracy of 92.13% on test data.

III. POS TAGGING APPROACHES

There are three types for POS tagging approaches called Rule based, Empirical based and Hybrid based. In Rule – based tagging, the hand written rules are used. Empirical POS taggers are further divided into Stochastic based taggers which either HMM based or Maximum Entropy models. There are two types of stochastic taggers Supervised and Unsupervised taggers.

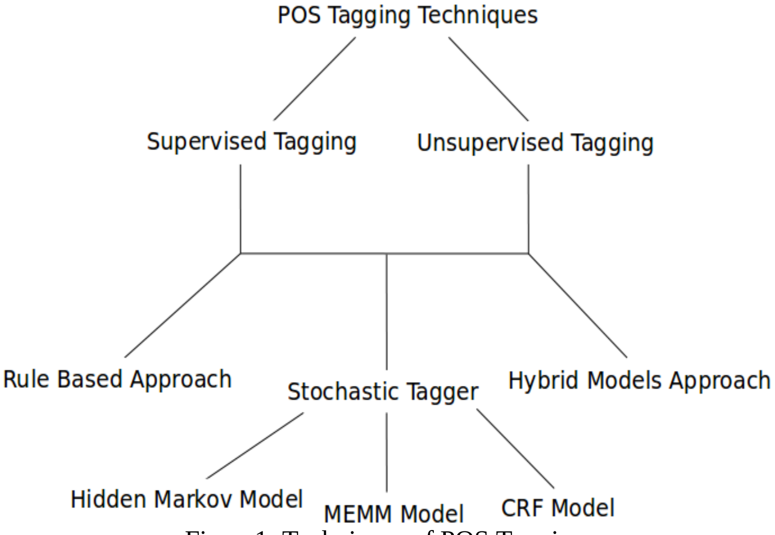

Figure1: Techniques of POS Tagging

A. **Rule Based Approach: - This** method uses the hand written rules and expert linguistic knowledge to assign appropriate POS tags to words in training data. Good experience and grammatical knowledge is required to obtain best results with use of this method. The rules used in this method are also called context frame rules. The acquisition cost of it is high. A widely used English POS-tagger is Brill"s tagger" based on rule-based approach [9].

B. **Empirical Based POS tagging Approach: -** The type of Empirical approach of parts of speech tagging is Stochastic based approach.

   a. ***Stochastic based POS tagging: -*** The Stochastic approach is beneficial to find out the most frequently used tag for a specific word in the annotated training data. After this, the same information is used to tag that word in the unannotated text. In stochastic approach various methods are used like N-grams, Maximum-Likelihood Estimation (MLE) or Hidden Markov Models (HMM). A large sized training corpus is required for stochastic approach. Two types of stochastic approach are [2]:

   b. ***Supervised models: -*** In Supervised POS Tagging for extracting information about the tagset, rule sets, word tag a pre- annotated corpus is required. For this approach if the corpus will be large then the results of evaluation will also be better. Examples for supervised POS taggers are:

   c. ***Hidden Markov Model (HMM) based POS tagging: -*** It calculates the probability of a given sequence of tags. According to the probability it specifies the most suitable tag for a word or token of a sentence that it occurs with the n previous tags, where the value of n is set to 1, 2 or 3 for practical purposes. The most useful algorithm for implementing an n-gram approach is HMM"s Viterbi Algorithm for tagging new text.

   d. ***Support Vector Machines Approach: -*** SVM is a machine learning algorithm has been applied to various practical problems like NLP. For dealing with all the requirements of modern NLP technology the SVM Approach is used because of combining simplicity, flexibility, robustness, portability and efficiency.

   e. ***Maximum Entropy Markov Model: -*** MaxEnt stands for Maximum Entropy Markov Model (MEMM). It is a conditional probabilistic sequence model. This method is used to represent multiple features of a word and also to handle long term dependency. It is based on the principle of maximum entropy which states that the least biased model is the one which maximizes entropy on the bases of all known facts. Each source state has an exponential model that takes the observation feature as input and provides output in form of distribution over possible next state. Output labels are associated with states. [9]

f. ***Conditional Random Field Model:*** - CRF stands for Conditional Random Field. It is a type of discriminative probabilistic model. It has all the advantages of MEMMs but avoids the label bias problem. CRFs are undirected graphical models (also known as random field) which are used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes [5].

g. ***Transformation-based POS tagging Approach:*** **-** In general, a large sized of pre-annotated corpus is required in supervised tagging approach But, Transformation –based tagging do not requires any pre-annotated corpus. In this Approach an untagged text is run through a tagging model to generate initial output. This is one approach for automatic rule induction after getting the output error correction is done. This way the taggers learn the correction rules by comparing the two sets of data. This process is repeated a number of times to achieve best results [2].

C. **Hybrid Models: -** Hybrid models are basically combination of rules based and statistical models. In Hybrid system, approach uses the combination of both rule-based and ML technique and makes new methods using strongest points from each method. It is making use of essential feature from ML approaches and uses the rules to make it more efficient. In this method, assign the most probable tag to the word using statistical after that, if wrong tag is found then by applying some rules tagger tries to change it [9][10].

IV. TAGSET

A tag set consist of tags that are used to represent the grammatical information of the language. The number of tags that are used for a language depends upon the information that user want to represent using a tag. A tagset can be too large according to requirement of researcher. For representing the context of words in a sentence of training data various tags are used if a word is acting as a noun then() NN tag is used like this for Pronoun (PRP) tag , Verb (V), Adjective (JJ), Conjunction (CC) can be used. For Punjabi Language Two POS tagger has been developed and both the taggers contain same tag set. A new tagset for Punjabi language is suggested by TDIL (Technical Development of Indian Languages) is used. TDIL proposed 36 pos tags for Punjabi language [10].

4.1 Tokenization and Speech Correction

Word Segmentation which is also known as tokenization focused on recognizing word boundaries exploiting orthographic word boundary delimiters, punctuation marks, written forms of alphabet and affixes [12]. The inherent difficulty associated with processing a variety of electronic text can cause the expansion of the complexity of the source code forming the tokenization program [13].

V. FEATURES FOR POS TAGGING

The Following features have been found to be very useful in POS tagging:

Suffixes: The next word of Current token is used as feature.

Prefixes: The previous word of Current token is used as feature.

Context Pattern based Features Context patterns are helpful for POS tagging. Eg.. Word prefixes and suffix context patterns.

Word length: Length of particular word is useful feature.

Static Word Feature: The previous and next words of a particular word are used as features. **Presence of Special characters:** Presence Special characters surrounding the current word are used as features.

VI. CONCLUSION

In this paper, a survey on developments of different POS tagger systems for Indian languages has been presented and also tried to give brief idea about the existing techniques that has been used to develop POS tagger for various Indian languages. The study shows that different approaches of POS tagging have been used which have performed very well and provided good results. But the most challenging task in this research field is to generate most efficient POS tagger for large training corpus which can give the best performance for different languages. In future we will try to evaluate the performance of POS tagger for Punjabi Language using other features, hybrid approach and we expect it will increase overall performance of the system.

REFERENCES

[1] Agrawal H, M. A. (2006). Part of speech tagging and chunking with conditional random fields. Proceedings of NWAI workshop.

[2] Antony P. J, S. K. P. (2011). Parts of speech tagging for Indian languages a literature survey. 34, 22–29.

[3] Garg N, P. S., Goyal V. (2012). Rule based hindi part of speech tagger. COLING, 163–174.

[4] Joshi N, M. I., Darbari H. (2013). Hmm based pos tagger for hindi. International Conference on Artificial Intelligence Soft Computing.

[5] Kaur M, S. S. K., Aggerwal M. (2014). Improving Punjabi part of speech tagger by using reduced tag set. International Journal of Computer Applications and Information Technology, 7, 142.

[6] Kumar D., J. G. S. (2010). Part of speech taggers for morphologically rich Indian languages: a survey. International Journal of Computer Applications, 6, 32–41.

[7] M, H. F. & UzZaman N, K. M. (2007). Comparison of different pos tagging techniques for bangla. InAdvances and Innovations in Systems, Computing Sciences and Software Springer Netherlands, 121–126.

[8] Mahar J A, M. G. Q. (2010). Rule based part of speech tagging of sindhi language. Signal Acquisition and Processing, 2010. ICSAP10. International Conference onIEEE, 101–106.

[9] Mohnot K, B. N. & Singh S P, K. A. (2014). Hybrid approach for part of speech tagger for hindi language. International Journal of Computer Technology and Electronics Engineering, 4, 25–30.