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

A Survey on Parts of Speech Tagging for Indian Languages

Jagjeet Singh* North West Institute of Engineering &Technology, Moga, India
Lakhvir Singh Garcha North West Institute of Engineering &Technology, Moga, India
Satinderpal Singh Sri Guru Granth Sahib World University, Fatehgarh Sahib, India

Abstract— Part of speech (POS) tagging is basically the process of automatically assigning its lexical category to each word according to its context and definition. Each word of sentence is marked in croups as corresponding to a particular part of speech like noun, verb, adjective and adverb. POS serves as a first step in natural language process applications like information extraction, parsing, and word sense disambiguation etc. this paper presents a survey on Part of Speech taggers used for Indian languages. The main problem of tagging is to find proper way to tag each word according to particular part of speech. Very less work has been done for POS tagging on Indian languages mainly due to morphologically richness. In this paper, various techniques are discussed that are used for development of POS tagger.

Keywords— POS, Natural Language Processing, HMM, MEMM

I. INTRODUCTION

With the advancement of technology, the demand of Natural Language Processing (NLP) is also increasing and it becomes very important to find out correct information from collection of huge data only on the basis of queries and keywords. Sometimes user tries to search data with help of query and get unimportant or irrelevant data instead of correct data. Due to complex structural effect, this problem occurs mostly with Indian languages as compared to others. To avoid this problem, POS tagging is the best application of NLP that assigns exact part of speech to each word of a text (Mohnot, K, 2014). It is the process of marking up a word in a corpus as corresponding to a particular part of speech use its definition, as well as its relation. POS tags are also known as word classes, morphological classes, or lexical tags to choose correct grammatical tag for word on the basis of linguistic feature. There are a number of approaches to implement part of speech tagger, i.e. Rule Based approach, Statistical approach and Hybrid approach. Rule-based tagger use linguistic rules to assign the correct tags to the words in the sentence or file. Statistical Part of Speech tagger is based on the probabilities of occurrences of words for a particular tag. Hybrid based Part of Speech tagger is combination of Rule based approach and Statistical approach. Part of Speech tagging is an important application of natural language processing. It is used in several Natural Languages processing based software implementation. Accuracy of all NLP tasks like grammar checker, phrase chunker, machine translation etc. depends upon the accuracy of the Part of Speech tagger. Tagger plays an important role in speech recognition, natural language parsing and information retrieval (Mehta, D. N, 2015). The paper presents a detail survey of various part of speech tagging techniques. Related work and past literature is discussed in section 2. Basic working of POS tagger is discussed in section 3. Type of POS tagging techniques and comparison based on different criteria is discussed in section 4. Finally, section 5 concludes the paper.

II. LITERATURE SURVEY

There have been many implementation of part of speech tagger using various approaches mainly for morphological rich languages. This section describes the work carried out by various researchers in this field.

In 2014, Pallavi Bagul et al. proposed a rule based pos tagger for Marathi language. The input sentence sent to tokenized function, the one which tokenizes the string into tokens and then comparing tokens with the Word Net. Tagging module assigned a tag to tokenized word and search for ambiguous word and pronoun. The ambiguous words were those words which can act as a noun and adjective in certain context, or act as an adjective and adverb in certain context. Then their ambiguity is resolved using Marathi grammar rules. Author used a corpus which is based on tourism domain called annotated corpus and 3 grammar rules are used for the experiment to resolve ambiguous word which acts a noun and adjective in certain context, or act as an adjective and adverb in certain context. In 2014, H.B. Patil et.al proposed a Partof-Speech Tagger for Marathi Language using Limited Training Corpora. It is also a rule based technique. Here sentence taken as an input generated tokens. Once token generated apply the stemming process to remove all possible affix and reduce the word to stem. SRR used to convert stem word to root word. They developed 25 SRR rule. The root-words that are identified are then given to morphological analyzer. The morphological analysis is carried out by dictionary lookup and morpheme analysis rules. Disambiguation is removed by the use of rule-base model or Hidden Markov Model. Based on the corpus they have identified 11 disambiguation rules that are used to remove the ambiguity. Stemming process removes all possible affixes, it change the meaning of stem word like (Anischit-Nischit).The size of the corpus is increased then more Rules can be discovered which will help to reduce the error rate.

In 2013, Jyoti Singh Nisheeth & Joshi Iti Mathur Proposed a Development of Marathi Part of Speech Tagger Using Statistical Approach. They used statistical tagger using Unigram, Bigram, Trigram and HMM Methods. To achieve higher accuracy they use set of Hand coded rules, it include frequency and probability. They train and test their model by calculating frequency and probability of words of given corpus. In unigram technique find out how many time each word occur in corpus and assign each word to most common tag. Bigram tagger makes tag suggestion based on preceding tag i.e. it take two tags previous and current tag. In Trigram provides the transition between the tags and helps to capture the context of the sentence. The probability of a sequence is just the product of conditional probabilities of its trigrams. Basic idea of HMM is assigns the best tag to a word by calculating the forward and backward probabilities of tags along with the sequence provided as an input. Powerful feature of HMM is context description. The POS taggers described here is very simple and efficient for automatic tagging, but it is difficult for Marathi as it is morphological rich language. In 2011, Nidhi Mishra & Amit Mishra proposed Part of Speech Tagging for Hindi Corpus. The system scans the Hindi (Unicode) corpus and then extracts the Sentences and words from the given Hindi corpus. Finally Display the tag of each Hindi word like noun tag, adjective tag, number tag, verb tag etc. and search tag pattern from database. The proposed model for Hindi language is apprehensible, but need to training data to increase accuracy. The efficiency of system judge on the basis of parameter of used need.

In 2012, Namrata Tapaswi Suresh Jain proposed a Treebank Based Deep Grammar Acquisition and Part-Of-Speech Tagging for Sanskrit Sentences. In the Sanskrit morphology meaning of the word is remain same. When affixes are added to the stem, words are differentiated at data base level directly. The input is one sentence per line, split the sentence in to words called lexeme. Read each word to find longest suffix, and eliminated the suffix until the word length is 2. Apply the lexical rules and assign the tag. Remove the disambiguity using context sensitive rules. For experimental result Author taken set of 100 words and manually evaluated, The system gives 90% correct tags for each word. The evaluation was done in two stages. Firstly by applying the lexical rules and secondly, after applying the context sensitive rule. The POS taggers described here is very efficient for Sanskrit but it is difficult for Marathi as affix is attached to root word so the meaning of word get change. Javed Ahmed MAHAR,

Ghulam Qadir MEMON, in 2010, proposed a system for "Rule Based Part of Speech Tagging of Sindhi Language". Take input text, and generate token. Once token generated search and compare selected word from lexicon (SWL) .If word is found one or more times, then store associated tag and if not found add that word into lexicon by generating linguistic rule for new word. The tagset contains 67 tags. A lexicon named SWL is developed having entries of 26366 words. Author also found the frequency for tag. For this purpose, set of 186 disambiguation rules are used for SPOS tagging system. The contextual information is used for rule-based approach and manually assigns a part of speech tag to a word. Accuracy of 96.28% was achieved from SPOS. When more words will be tagged and rules will be added then accuracy will be increased. In 2012, Kamal Sarkar, Vivekananda Gayen proposed "A Practical Part-of-Speech Tagger for Bengali". The system has two major phases: training phase and testing phase. In the training phase, the system is trained on a handful of POS tagged Bengali sentences by computing tag transition probabilities and word likelihoods or observation probabilities. In the testing phase, untagged Bengali sentences are submitted to the system for tagging. Viterbi algorithm is used for finding the most likely tag sequence for each sentence in the input document. Author implemented a supervised Bengali trigram POS Tagger from the scratch using a statistical machine learning technique that uses the second order Hidden Markov Model (HMM).The performance of the POS tagger can be improved by introducing more accurate method for unknown word handling.

III. POS TAGGER

The broad utilization of internet for making search of information is difficult due to the search systems consist container of words which causes problem in retrieval due to synonyms. There is need to accept the word boundary between what kinds of query information are submitted by humans and what kinds further result get (Tapaswi, N., & Jain, S., 2012). So for text indexing and retrieval uses POS information. POS tagging is used as an early stage of text analysis in many applications such as subcategory acquisition, text to speech synthesis and alignment of parallel corpora. POS tagging is a necessary pre-module and building block for various NLP tasks like Machine translation, Natural language text processing and summarization, User interfaces, Multilingual and cross language information retrieval, Speech recognition, Artificial intelligence, Parsing , Expert system and so on . Parts of speech (POS) tagging are one of the most well studied problems in the field of Natural Language Processing (NLP). Different approaches have already been tried to automate the task for English and other western languages there are large numbers of POS tagger available for English language which has got satisfactory performance but cannot be applied to Marathi language. Part-of-speech tagging in Marathi language is a very complex task as Marathi is highly inflectional in nature & free word order language (Mehta, D. N., & Desai, N. P. 2015). The process of assigning description to the given word is called Tagging. The descriptor is called tag. The tag may indicate one of the parts-of-speech like noun, pronoun, verb, adjective, adverb, preposition, conjunction, and interjection. The input (Raw Text) is tokenized and a corpus is used for detecting the corresponding part of speech of each token in the sentence. For correct POS tagging, training the tagger, corpus and a proper tagset is also important Disambiguation is the most difficult problem in tagging. The ambiguity which is identified in the tagging module is resolved using the grammar rules.

A. Architecture of POS tagger

1) ***Tokenization:*** Tokenization is the process of separating tokens from raw text. Words are separated by white spaces or punctuation marks. The sentence is segmented by using white space because the occurrence of white

space indicates the existence of a word boundary. There are various morphological problems where this approach fails. So by using this we can easily find out the tokens from the sentence. The given text is divided into tokens so that they can be used for further analysis. The tokens may be words, punctuation marks, and utterance boundaries (Bagul, P. et al., 2014).

2) ***Ambiguity look-up:*** This is to use lexicon and a guesser for unknown words. While lexicon provides list of word forms and their likely parts of speech, guessers analyze unknown tokens. Compiler or interpreter, lexicon and guesser make what is known as lexical analyser [11].
3) ***Ambiguity Resolution:*** This is also called disambiguation. Disambiguation is based on information about word such as the probability of the word. Disambiguation is also based on related information or word/tag sequences. For example, the model might prefer noun analyses over verb analyses if the preceding word is a preposition or article [11]. Disambiguation is the most difficult problem in tagging. The ambiguity which is identified in the tagging module is resolved using the Marathi grammar rules.
4) ***WordNet:*** The main relation among words in WordNet is synonymy. WordNet is an electronic database which contains parts of speech of all the words which are stored in it. It is trained from the corpus for higher performance and efficiency (Bagul, P. et al., 2014). WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The majority of the WordNet's relations connect words from the same part of speech (POS). Thus, WordNet really consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers. Cross-POS relations include the "morph semantic" links that hold among semantically similar words sharing a stem with the same meaning [10].
5) ***Corpus:*** For correct POS tagging, training the tagger well is very important, which requires the use of well annotated corpora. Annotation of corpora can be done at various levels which include POS, phrase or clause level, dependency level etc. Corpus linguistics is the study of language as expressed in samples (corpora) of "real world" text. Corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. The plural form of corpus is corpora. Some popular corpora are British National Corpus (BNC), COBUILD/Birmingham Corpus, and IBM/Lancaster Spoken English Corpus (Bagul, P. et al., 2014).
6) ***Tagset:*** Apart from corpora, a well-chosen tagset is also important. The language tagset represents parts of speech and consist on syntactic classes. According to contextual and morphological structure, natural languages are different from each other.In the top level the following categories are identified as universal categories for all ILs and hence these are obligatory for any tagset. Some common tags: [N] Nouns [V] Verbs, [PR] Pronouns, [JJ] Adjectives, [RB] Adverbs, [PP] Postpositions, [PL] Participles, [QT] Quantifiers, [RP] Particles, [PU] Punctuations (Mahar, J. A., & Memon, G. Q., 2010).

IV. POS TAGGING TECHNIQUES

The POS tagger can be implemented by using either a supervised technique or an unsupervised technique. Supervised POS taggers are based on pre-tagged corpora [6], which are used for training to learn information about the word-tag frequencies, rule and tag set, sets etc. The performance of the models generally increases with the increase in size of these corpora.

Unsupervised POS tagging models do not require pretagged corpora. Instead, they use those methods through which automatically tags are assigned to words [6]. Advanced computational methods like the Baum-Welch algorithm to automatically include tag sets, transformation rules etc. Under these two categories different approaches have been used for the implementation of POS taggers such as:

A. Rule Based Approach / Transformation Based:

The rule based POS tagging approach that uses a set of hand written rules. Rule base taggers depend on word list or lexicon or dictionary to assign appropriate tag to each word. The tagger divided into two stages. First, it search words in dictionary and second, it assigns a tag by removing disambiguity of words using linguistic features of word.

On the basis of level rule divided as lexical rules act in a word level, each sentence splits into small words called lexeme or token And, the context sensitive rules act in a sentence level, to check the grammar for the sentence. The transformation based approach is similar to the rule based approach in the sense that it depends on a set of rules for tagging. The transformation based approaches use a pre-defined set of handcrafted rules as well as automatically induced rules that are generated during training (Bagul, P. et al., 2014)..

The main drawback of rule based system is that it fails when the text is not present in lexicon. Therefore the rule based system cannot predict the appropriate tags.

B. Statistical Approach / Stochastic Tagger:-

A stochastic approach assign a tag to word using i frequency, probability or statistics. From the annotated training data it "selects the most likely tag for the word" and uses same information to tag that word in the unannotated text (Bagul, P. et al., 2014). Stochastic tagger as a simple generalization of the stochastic taggers generally resolves the ambiguity by computing the probability of a given word (or the tag).

The drawbacks of this approach are that it can come up with sequences of tags for sentences that are not acceptable according to the grammar rules. So, it determines the best tag for a word by calculating the probability of previous tags on n value, where the value of n is set to 1, 2 or 3 are known as the Unigram, Bigram and Trigram models.

1) ***Hidden Markov Model:*** **-** HMM stands for Hidden Markov Model. HMM is a generative model. The model assigns the joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets .It is advantageous as its basic theory is elegant and easy to understand. Hence it is easier to implement and analyze. It uses only positive data, so they can be easily scaled. It has few disadvantages. In order to define joint probability over observation and label sequence HMM needs to enumerate all possible observation sequence. Hence it makes various assumptions about data like Markovian assumption i.e. current label depends only on the previous label. Also it is not practical to represent multiple overlapping features and long term dependencies. Number of parameter to be evaluated is huge. So it needs a large data set for training (Bagul, P. et al., 2014).
2) ***Maximum Entropy Markov Model:*** -MaxEnt stands for Maximum Entropy Markov Model (MEMM). It is a conditional probabilistic sequence model. It can represent multiple features of a word and can also handle long term dependency. It is based on the principle of maximum entropy which states that the least biased model which considers all known facts is the one which maximizes entropy. Each source state has an exponential model that takes the observation feature as input and output a distribution over possible next state. Output labels are associated with states. [10] The large dependency problem of HMM is resolved by this model. Also, it has higher recall and precision as compared to HMM. The disadvantage of this approach is the label bias problem. The probabilities of transition from a particular state must sum to one. MEMM favors those states through which less number of transitions occurs (Mohnot K & Singh S P, 2014).
3) ***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 without the label bias problem. CRFs are undirected graphical models (also know as random field) which is used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes. [11]
4) ***Hybrid Approach:*** This approach combines the advantages of both of the above approaches namely rule based approach and stochastic approach. Words in this technique are first tagged probabilistically and then as post processing, linguistic rules are applied to tag tokens. Accuracy of taggers based on this approach generally gives good results than other techniques (Mehta, D. N, 2015).
5) ***Neural Tagger****:* Neural taggers are based on neural networks which learn the parameters of POS tagger from a representative training data set. The MLP-tagger is trained with error back-propagation learning algorithm. The performance has shown better than stochastic method (Raju, S. B., 2002).

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. Eample. 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

This paper presents the study of various POS tagging approaches and the approaches used to develop POS tagger for various Indian languages. The study shows that POS tagger plays an important role to create NLP based applications but the problem is to assign exact POS tagger to each word of sentence due to morphological richness of languages. However many researchers worked in this field but its still challenge to achieve best results. In future we will try to evaluate the performance of POS tagger for Hindi Language using Apache open NLP that supports many NLP applications and we expect it will increase overall performance of the system.

REFERENCES

[1] Mohnot, K., Bansal, N., Singh, S. P., & Kumar, A.(2014). Hybrid approach for Part of Speech Tagger for Hindi language. *International Journal of Computer Technology and Electronics Engineering*, 4(1), 25-30.

[2] Mehta, D. N., & Desai, N. P. (2015). A Survey on Part-Of-Speech Tagging of Indian Languages. *History*, *43*(198), 125-131.

[3] Raju, S. B., Chandrasekhar, P. V., & Prasad, M. K. (2002, December). Application of multilayer perceptron network for tagging parts-of-speech. In*Language Engineering Conference, 2002. Proceedings* (pp. 57-63). IEEE.

[4] Bagul, P., Mishra, A., Mahajan, P., Kulkarni, M., & Dhopavkar, G. (2014). Rule Based POS Tagger for Marathi Text. *Proc. Int. J. Comput. Sci. Inf. Technol.(IJCSIT)*, *5*(2), 1322-1326.

[5] Patil, H. B., Patil, A. S., & Pawar, B. V. (2014). Part-of-Speech Tagger for Marathi Language using Limited Training Corpora. *International Journal of Computer Applications*, 0975-8887.

[6] Singh, J., Joshi, N., & Mathur, I. (2013, August). Development of Marathi part of speech tagger using statistical approach. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*(pp. 1554-1559). IEEE.

[7] Mishra, N., & Mishra, A. (2011, June). Part of Speech Tagging for Hindi Corpus. In *Communication Systems and Network Technologies (CSNT), 2011 International Conference on* (pp. 554-558). IEEE.

[8] Tapaswi, N., & Jain, S. (2012, September). Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit sentences. In *Software Engineering (CONSEG), 2012 CSI Sixth International Conference on* (pp. 1-4). IEEE.

[9] Mahar, J. A., & Memon, G. Q. (2010, February). Rule Based Part of Speech Tagging of Sindhi Language. In *Signal Acquisition and Processing, 2010. ICSAP'10. International Conference on* (pp. 101-106). IEEE.

[10] Sarkar, K., & Gayen, V. (2012, November). A practical part-of-speech tagger for Bengali. In *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on* (pp. 36-40). IEEE.

[11] http://wordnet.princeton.edu/

[12] http://language.worldofcomputing.net/pos-tagging/parts-of-speechtagging.htm

