ISSN:0975-9646

Survey of various POS tagging techniques for Indian regional languages

Shubhangi Rathod [#1], Sharvari Govilkar [*2]

[#1,2]*Department of Computer Engineering,*
University of Mumbai, PIIT, New Panvel, India

***Abstract*—Part of Speech tagging (POS) is an important tool for processing natural languages. It is one of the simplest as well as most stable and statistical model for many Natural language processing (NLP) applications. It is the process of marking up a word in a corpus as corresponding to a particular part of speech like noun, verb, adjective and adverb. There are many challenges in POS tagging like Foreign words, Ambiguities, ungrammatical input etc. In this paper, comparison of various POS tagging techniques for Indian regional languages has been discussed elaborately.**

Keywords— POS-Part of Speech, Rule Based Approach Statistical Approach, Hybrid Approach.

I. INTRODUCTION

Natural language processing is the skill of a computer program to understand human language as it is spoken. It is a component of computer science, linguistics and artificial intelligence. To build NLP application is a difficult because human speech is not always specific. NLP is a process of developing a system that can read text and translate between one human language and another.

The work on Part-of-Speech (POS) tagging has begun in the early 1960s [2]. Part of Speech tagging is an important tool for processing natural languages. It is one of the simplest as well as most stable and statistical model for many NLP applications POS Tagging is an initial stage of information extraction, summarization, retrieval, machine translation, speech conversion [2].

The part of speech tagging is a process of assign appropriate parts of speech tags like noun, verb, adjective and adverb to each word in a input sentence. 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 [3]. The main challenge in POS tagging is to resolving the ambiguity in possible POS tags for a word.

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

The process of POS tagging consists of these stages Tokenization, Assign a tag to tokenized word and search for Ambiguous word. Disambiguation is done by analyzing the linguistic feature of the word, its preceding word, its following word, etc. Considerable work is already done for foreign languages if we look at the same scenario for South-Asian languages such as Marathi and Hindi, it find out that not much work has been done. As Marathi is a morphological rich language and unavailability of annotated corpora.

In 2014, Pallavi Bagul, Archana Mishra, Prachi Mahajan, Medinee Kulkarni, Gauri Dhopavkar [1] 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 assigns a tag to tokenized word and search for ambiguous word and pronoun. The ambiguous words are 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.

H.B. Patil, A.S. Patil, B.V. Pawar [2] proposed a Part-of-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.

Jyoti Singh Nisheeth Joshi Iti Mathur [3] 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.

Nidhi Mishra, Amit Mishra [4] 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.

Namrata Tapaswi Suresh Jain [5] 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 [6], 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.

Kamal Sarkar, Vivekananda Gayen [7] 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 [5]. 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 [2]. 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 [2].

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 [1] [11].

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].

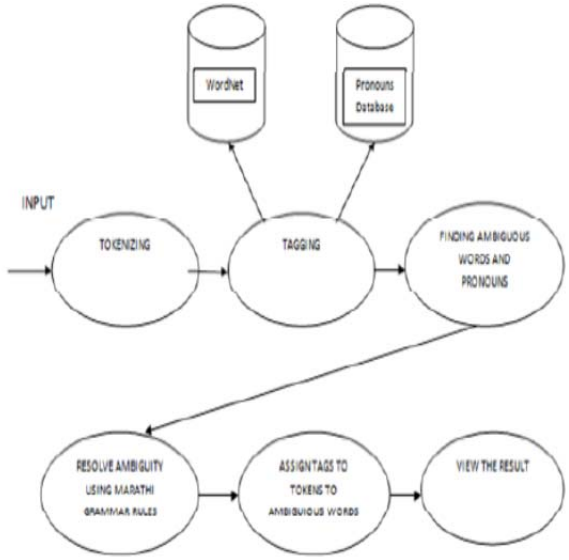

Fig. 1 Process Overview [1]

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 [1]. 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 [1]. 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, IBM/Lancaster Spoken English Corpus.

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 [6].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.

IV. POS TAGGING TECHNIQUES

The POS tagger can be implemented by using either a supervised technique or an unsupervised technique.

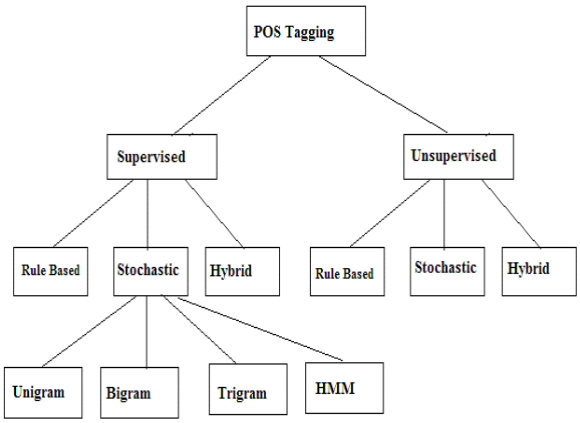

Fig. 2 Classification of POS tagging models

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 pre-tagged 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 [6].

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 [5]. 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 [8]. 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 un-annotated text [1] [5]. 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 is 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 [5,8]. Hybrid approach

Metaio The hybrid approach is a combination of Rule based approach and statistical approach, that assign most probable tag to the word using statistical after that, if disambiguity is found then by applying grammar rules tagger tries to change it. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

1) Unigram: It consider one word at a time and assigns each word to its most common tag.P (ti/wi) = freq (wi/ti)/freq (wi) .Here Probability of tag for current word is calculated by frequency count of word given tag divided by frequency count of that particular word[3].

2) Bigram: It consider two tag the preceding tag and current tag into account.P (ti/wi) = P (wi/ti). P (ti/ti-1).Here P (wi/ti) is the probability of current word given current tag.P (ti/ti-1) is the probability of a current tag given the previous tag [3].

3) *Trigram:* A model based approach uses prior knowledge of 3D objects in the environment along with their appearance [14]. It use current tag and based on previous two tags.P (ti/wi) = P (wi/ti). P (ti/ti-2, ti-1)Where ti and wi indicate tag sequence and word sequence respectively. P (wi/ti) is the probability of current word given current tag. Here, P(ti|ti-2, ti-1)is the probability of a current tag given the previous two tags [3].

4) *Hidden Markov Model (HMM):* It is called Hidden Markov model because We cannot determine the exact sequence of tags that generated and calculate usingt = argmax P(w, t) [8] and it is based on the Markovian assumption that the current tag depends only on the previous n tags.The HMM use a transition probability(i.e. forward tag and backward tags) to assign a tag.P (ti/wi) = P (ti/ti-1). P (ti+1/ti). P (wi/ti)P (ti/ti-1) is the probability of current tag given previous tag. P(ti+1/ti) is the probability of future tag given current tag. P (wi/ti) Probability of word given current tag [3].

TABLE I. COMPARISON OF POS TAGGING TECHNIQUES

Techniques	**Description**	**Advantages**	**Disadvantages**	**Accuracy**
Rule Based	Uses a set of hand written rules.	1) Small set of simple rules. 2) Less stored information	Generally less accurate as compared to stochastic taggers.	marathi-78.82% Sindhi-96.28% Sanskrit-90%
Stochastic	Probabilistic depending on the N previous tags (1, 2, and 3) called unigram, bigram or trigram frequencies in a training corpus. HMM use transition probability.	Generally more accurate as compared to rule based taggers	Relatively complex. Require vast amounts of stored information	Marathi unigram, Bigram, Trigram and HMM gives the accuracy of 77.38%, 90.30%, 91.46% and 93.82% respectively. Bengali bigram tagger-74.33 and trigram - 78.68
Hybrid	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.	Having higher accuracy than individual rule based or statistical approach	Not assign correct tag to an unknown word	Hindi - 79.66% Bengali- 95%

V. CONCLUSION

Automatic POS tagging makes errors because many high frequency words of part-of-speech are ambiguous. Rule-based tagging assigns a word all possible tags and the uses context rules to disambiguate. Statistical tagging assigns a word its most likely tag, based on the n-set values frequencies in a training corpus. Hybrid-based tagging combines the two approaches.

ACKNOWLEDGMENT

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

REFERENCES

[1] Pallavi Bagul, Archana Mishra, Prachi Mahajan, Medinee Kulkarni, Gauri Dhopavkar, "Rule Based POS Tagger for Marathi Text" 2014 in proceeding of: International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1322-1326.

[2] H.B. Patil, A.S. Patil, B.V. Pawar "Part-of-Speech Tagger for Marathi Language using Limited Training Corpora" 2014 in International Journal of Computer Applications (0975 – 8887) Recent Advances in Information Technology.

[3] Jyoti Singh Nisheeth Joshi Iti Mathur" Development of Marathi Part of Speech Tagger Using Statistical Approach" , Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on.

[4] Nidhi Mishra, Amit Mishra,"Part of Speech Tagging for Hindi Corpus" 2011 in proceeding of : International Conference on Communication Systems and Network Technologies

[5] Namrata Tapaswi Suresh Jain , "Treebank Based Deep Grammar Acquisition and Part-Of-Speech Tagging for Sanskrit Sentences" Software Engineering (CONSEG), 2012 CSI Sixth International Conference on.

[6] Javed Ahmed MAHAR Ghulam Qadir MEMON,"Rule Based Part of Speech Tagging of Sindhi Language " 2010 proceeding of International Conference on Signal Acquisition and Processing.

[7] Kamal Sarkar , Vivekananda Gayen "A Practical Part-of-Speech Tagger for Bengali".2012 in proceeding of Third International Conference on Emerging Applications of Information Technology (EAIT)

[8] Fahim Muhammad Hasan "Comparison Of Different Pos Tagging Techniques For Some South Asian Languages".

[9] Sankaran Baskaran1,Kalika Bali1, Tanmoy Bhattacharya2, Pushpak Bhattacharyya3, Monojit Choudhury1,Girish Nath Jha4, Rajendran S.5, Saravanan K.1, Sobha L.6, and KVS Subbarao "A Common Parts-of-Speech Tagset Framework for Indian Languages ".

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

[11] http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.htm