A Survey on Parts of Speech Tagging for Indian Languages

Neetu Aggarwal
BGIET, Sangrur

Amandeep kaur Randhawa
BGIET,Sangrur

ABSTRACT
This paper describes the survey on POS (Part of Speech) tagging for various Indian Languages. Various approaches concerned for POS tagging of sentences written in Indian languages are discussed in this paper . Indian Languages have rich morphological effect so a no. of problems occur while tagging the sentences written in various languages. A lot of POS tagging work has been done by the researchers for various languages using different approaches HMM(Hidden Marcov Model) , SVM (Support Vector Machine) , ME (Maximum Entropy) etc.

Keywords
Natural Language Processing, Part of Speech tagging, Tagset, Indian Languages

1. INTRODUCTION
The main objective of Natural Language Processing is to facilitate the interaction between human and machine. POS tagging is the process of attaching the best grammar tag like to each word of a sentence of some language. A word in a sentence can act as a verb, noun , pronoun , adjective , adverb, conjunction , preposition etc so POS is defined as the grammatical information of each word of a sentence. While assigning a POS tag it is necessary to determine the context of the word i.e. whether it is acting like a noun, adjective, verb etc. Sometime a word can act as a noun in one sentence and in another sentence it can give the sense of verb. So before selecting a POS tag for a word the exact context of the word must be clear.

For Indian languages it is a difficult task to assign the correct POS tag to each word in a sentence because of some unknown words in Indian languages. The earlier work that has been done for Indian languages was based rule based approaches. But the rule-based approach needs proper language knowledge and hand written rule. 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. The set of tags assigned by a part of speech tagger may contain just a dozen tags so such a big tagset can arise the difficulty in the tagging process. POS tagging is helpful in various NLP tasks like Information Retrieval, Machine Translation , Information Extraction , Speech Recognition etc. For Indian languages researchers find difficulty in writing linguistic rules for rule based approaches because of morphological richness . The other main issue after morphological richness of Indian Languages is Ambiguity. It is very time consuming process to assign a POS tag to each word according to its context in sentence by hand and that is why POS Tagging is becoming a challenging problems for study in the field of NLP.

2. POS TAGGING APPROACHES
There are three categories for POS tagging approaches called Rule based, Empirical based and Hybrid based . In Rule – based tagging the rule that used are hand – written. Empirical POS taggers are further divided into Stochastic based taggers which either HMM based that use Decision Trees or Maximum Entropy models. There are two types of Stochastic taggers Supervised and Unsupervised taggers.

2.1 Rule Based Approach
In Rule-based approach handwritten rules and grammatical information is used to assign POS tags to words in training data. These rules are often known as context frame rules.

A widely used English POS-tagger is Brill's tagger" based on rule-based approach.

2.2 Empirical Based POS tagging Approach
The type of Empirical approach of parts of speech tagging is Stochastic based approach.

2.2.1 *Stochastic based POS tagging*
The Stochastic approach is helpful to find out the most frequently used tag for a specific word in the annotated training data and uses this information 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:

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:

Hidden Markov Model (HMM) based POS tagging:
It calculates the probability of a given sequence of tags. By calculating 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.

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.

Unsupervised models

As in Supervised POS tagging approach a pre-annotated training corpus is required, in unsupervised approach there is no requirement of a pre-annotated corpus. Instead, researchers use advanced computational techniques like the Baum-Welch algorithm to automatically induce tagsets, transformation rules, etc. Evaluation of the probabilistic information or build up the contextual rules needed by Rule based systems or Transformation based systems is performed for Stochastic Taggers.

2.3 Transformation-based POS tagging Approach

In general, a large sized of pre-annotated corpus is required in supervised tagging approach But in Transformation –based tagging a pre-annotated corpus is not required. 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. For obtaining the better performance. This process is repeated a no. of times.

3. TAGSET

A tag set consist of tags that are used to represent the grammatical information of the language. The number of tags that we use for a language depends upon the information that we 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 consist 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.

4. LITERATURE SURVEY FOR INDIAN LANGUAGES

Different approaches have been used for part-of speech tagging and different researchers have developed POS taggers for various languages Foreign Languages like English, Arabic and other European languages have more POS taggers than Indian languages. Indian Languages for which POS taggers have been developed are Hindi, Bengali, Panjabi and Tamil.

In this paper [1] 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 [2] 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 than it does 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 [3] A POS tagging for Punjabi language using Hidden Marcov Model has been used by Sapna Kanwar, Mr Ravishankar, Sanjeev Kumar Sharma and used a Bi-gram Hidden Markov Model to solve the part of speech tagging problem. During experimental results they note that the general HMM based method doesn't perform well due to data deficiency problem.

This paper [4] introduces A Machine learning algorithm for Gujarati Part of Speech Tagging has been used by Chirag Patel and Karthik Gali. The machine learning part is performed using a CRF model. The algorithm has achieved an accuracy of 92% for Gujarati texts where the training corpus is of 10,000 words and the test corpus is of 5,000 words. From the experiments they observed that if the language specific rules can be formulated in to features for CRF then the accuracy can be reached to very high extents.

In this paper [5] Sumeer Mittal used N Gram Model for Part of Speech Tagging of Punjabi Language. A Bi-gram Model has been used to solve the part of speech tagging problem. An annotated corpus was used for training and estimating of bi gram probabilities. During experimental results he noted that the general-Gram based method doesn't perform well due to unknown words (foreign language words or due to spelling mistakes) problem.

In this paper [6] Kavi Narayana Murthy and Srinivasu Badugu proposed a new approach to automatic tagging without requiring any machine learning algorithm or training data using a morphological analyzer and a fine-grained hierarchical tag-set.. They have worked on Telugu and Kannada languages. They argue that the critical information required for tagging comes more from word internal structure than from the context and they show how a well designed morphological analyzer can assign correct tags and disambiguate many cases of tag ambiguities too.

This paper [7] specfies A Comparison of Unigram, Bigram, HMM and Brill'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 Brill'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 [8] 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%.

In this paper [9] 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 [10] Adwait Ratnaparkhi used a Maximum Entropy Model for POS tagging. He presents a statistical model which trains from a corpus annotated with Part-Of - Speech tags and assigns them to previously unseen text with state-of-the-art accuracy (96.6%). The model can be classified as a *Maximum Entropy* model and simultaneously uses many contextual "features" to predict the POS tag. Furthermore, He demonstrates the use of specialized features to model difficult tagging decisions.

5. PROBLEMS OF PART OF SPEECH TAGGING

The main problem in part-of speech tagging is Ambiguity. It is possible that a word in a sentence can act as more than one meaning so it can have more than one tag so such situation arise the problem of Ambiguity. To solve this problem we consider the context instead of taking single word. For example-

auh ie`k imhnqI kuVI sI ausdy mW-bwp ny ausdw pUrw swQ id`qw qy auh swry ausdI sPlqw qy bhuq KuS sn[

In this example word '**ਉਹ**' is both acting as a singular pronoun and plural pronoun. Since word **ਉਹ** occur in between the sentence and also the word next to it is not a noun so it may be a pronoun. The previous word of the sentence determines the type of pronoun that is singular or plural. By looking at the context of the word the correct POS of a word in a sentence can be identify.

6. 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 prefix 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.

7. EVALUATION METRICES

The evaluation metrics for the data set is precision, recall and F-Measure. These are defined as following:-

Recall = Number of correct answer given by system / Total number of words.

Precision = Number of Correct answer / Total number of words.

F-Measure = Recall *Precision / Recall + Precision

8. CONCLUSION

In this paper work, we tried to give a brief idea about the existing approaches that have been used to develop POS tagger tools. We have presented a survey on developments of different POS tagger systems for Indian languages. We found out from the survey that for Indian Languages Rule-based, Supervised, Unsupervised , Transformation based POS tagging approaches have been used which have given good performance results. In each research work the most challenging task is to generate the 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 and we expect it will increase overall performance of the system

9. REFERENCES

[1] Antony P J, Vidyapeetham A.V and Dr. Soman K P . 2011 "Parts Of Speech Tagging for Indian Languages: A Literature Survey" International Journal of Computer Applications (0975 – 8887) Volume 34– No.8

[2] Aggarwal H, Anirudh Amni. 2006 "Part of Speech Tagging and Chunking with Conditional Random Fields" in the proceedings of NLPAI Contest

[3] Kanwar S., Ravishankar, Sharma S. K. 2011 "POS of Punjabi language by Hidden Marcov Model" An International Journal of Engineering Sciences ISSN: 2229-6913 Issue Vol. 1.

[4] Patel C. and Gali K .2008 "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields" Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 117–122

[5] Mittal.S, Sethi N.S. and Sharma S.K.. 2014 "Part of Speech Tagging of Punjabi Language using N Gram Model" International Journal of Computer Applications (0975 – 8887) Volume 100– No.19.

[6] Murthy K.N. and Badugu S. 2013A "New Approach to Tagging inIndian Languages" Research in Computing Science 70, pp. 45–56

[7] Hasan F.M., UzZaman N, Khan M. 2006 "Comparison of Different POS Tagging Techniques (n-grams, HMM and Brill's Tagger) for Bangla", International Conference on Systems, Computing Sciences and Software Engineering (SCS2 06) of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 06)

[8] Garg N. ,Goyal V.and Preet S.(2012) "Rule Based Hindi Part of Speech Tagger" Proceedings of COLING : Demonstration Papers, pages 163–174.

[9] Kaur M., Aggerwal M. and Sharma S. K.(2015) " Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set" International Journal of Computer Applications & Information Technology Vol. 7, Issue II (ISSN: 2278-7720).

[10] Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania, pp. 133–142.