International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 8
August 2017

Part-of-Speech Tagging of Hindi Language Using Hybrid Approach

Vijeta Khicha
Yagyavalkya Institute of Technology, Jaipur, India
Mantosh Manna
Yagyavalkya Institute of Technology, Jaipur, India

ABSTRACT—*This Natural language processing is used to understand natural languages by parsing the text. The analysis of natural language processing based on various intermediate tasks. These tasks don't need complete understanding and knowledge of the language. The Part-of–Speech (POS) Tagging is one of the processing tasks which tag every word in a text according to its presence in the given text. Many POS tagging approaches were developed based on supervised and unsupervised learning for understanding languages. Hindi is a rich morphological language. In this paper, a Hybrid Approach is build using Hidden Markov Model (HMM) and Rule Based Tagging on the Hindi language. HMM maximizes the likelihood and tag sequence probability and grammar rules are applied to increase the accuracy of tag words. The presented system uses a pre-tagged corpus of around 13,000 words and several tagset like noun, pronoun, verb, etc. The system yields 96.01% of average precision and 89.32% of average accuracy.*

KEYWORDS—Part-of-Speech Tagging, Hindi, Hidden Markov Model, Ruled Based Tagging.

INTRODUCTION

Corpus is "a body of language in a large and structured set of data and used as a key for natural language processing". Corpus is generally in the modes of written text, printed text or sample of spoken words or combination of all. In annotation process, input, as well as output, is natural languages like English, Hindi, etc. There are different levels of corpus annotation like Morphological analysis, POS tagging, Chunk tagging, etc. POS tagging is a basic step for language processing and can work as the first phase in other language processing tasks. The work on Part-of-Speech (POS) tagging for natural language tagging has begun in the early 1960s. For Indian languages researcher, it's difficult to write linguistic rules using rule based approaches because of morphological richness. POS Tagging is the process of marking up a token in a sentence as a particular POS tag or lexical belonging to a particular class (noun, adverb, pronoun, etc) based on its context in the sentence, its definition, and its morphological information. Formally it can be defined as, "Given a meaningful sequence of words $w_1 \ldots w_n$, the system has to assign respective POS tags $t_1 \ldots t_n$ to input sequence". Part of speech tags or morphological tags provides useful information about a word. They give relevance of word in a given context that is the role of the word in a given sentence. POS Tagging is a basic tool for various applications of NLP, such as Text recognition, Opinion mining, Named-entity recognition, Machine learning, etc. The Hindi language is a feature-rich language. It is very time consuming to tag each every word according to the context manually. So to remove this we have to use POS Tagging in which we use a rule to assign a tag to words. These techniques are divided into two broad categories supervised POS tagging and unsupervised tagging. Supervised POS tagging is based on the pre tagged corpus and unsupervised POS tagging don't need pre tagged data.

Both categories are further divided into 2 broad categories stochastic-based approach and rule based approach. The Hidden Markov Model is based on the supervised learning based stochastic-based approach. This model required data set and pre human made tables to tag new incoming data whereas Rule Based Tagging required deep knowledge of language and each and every rule is written based on grammar. In this paper, we have applied Hidden Markov model with rule based tagging approach for tagging data. The same word may have different meaning with respect to the context which causes ambiguity issue. So the most challenging objective in the area of POS Tagging for the Hindi language is identifying the ambiguities in tags

RELATED WORK

There are numerous of POS tagger available in many languages like English, Myanmar, Hindi, etc based on various techniques and approaches. The first corpus was built in 1991 contain one million words of English called ‘The Brown Corpus’ and achieved an accuracy of 95%[1].Zin and Thein et al.,2009 developed a system for efficient part-of-speech tagging for Myanmar language based on pre-tagged training data of 1,000,000 words and statistical approach using HMM. 97.56%of accuracy is achived[2]. Aniket Dalat et al.,2006 proposed a system using Maximum entropy Markov model for Hindi system based on a morphological and lexical feature of a language. The system contains 27 POS tags and corpus of 15562 with an accuracy of 94.81%[3].Smriti Singh et al.,2006 develop POS tagger based on decision tree base learning system has detailed linguistic analysis of language using 4-fold cross validation and obtained accuracy of 93.45%[4].Himanshu Aagarwal et al., 2006 developed a model using Conditional Random Filed for Hindi. Achieved an accuracy of 82.67% using 21000 words corpus [5] . Garg et al. (2012) developed a Rule-based system and obtained average precision of 85.47% on different datasets [6].

We can see that most POS tagger is using stochastic or rule based approach. In this paper, we are implementing Hybrid approach using Hidden Markov Model and rule based approach. In next sections, we describe the proposed approach.

SYSTEM DESCRIPTION

The proposed system used a hybrid approach (hidden Markov model and rule based model).The first system used Hidden Markov model using probabilistic analysis of pre-tagged corpus and then apply rule based model on remaining untagged corpus. The system has 32 tags set with 27 tags set from IIIT - Hyderabad tagset (POS tagger 2007) and 5 new special tag set have been added. Around 13,000 Hindi words pre-corpus tagged set is prepared in .xml format. The system is build using Java language.

The system is developed in various stages shown in fig.1

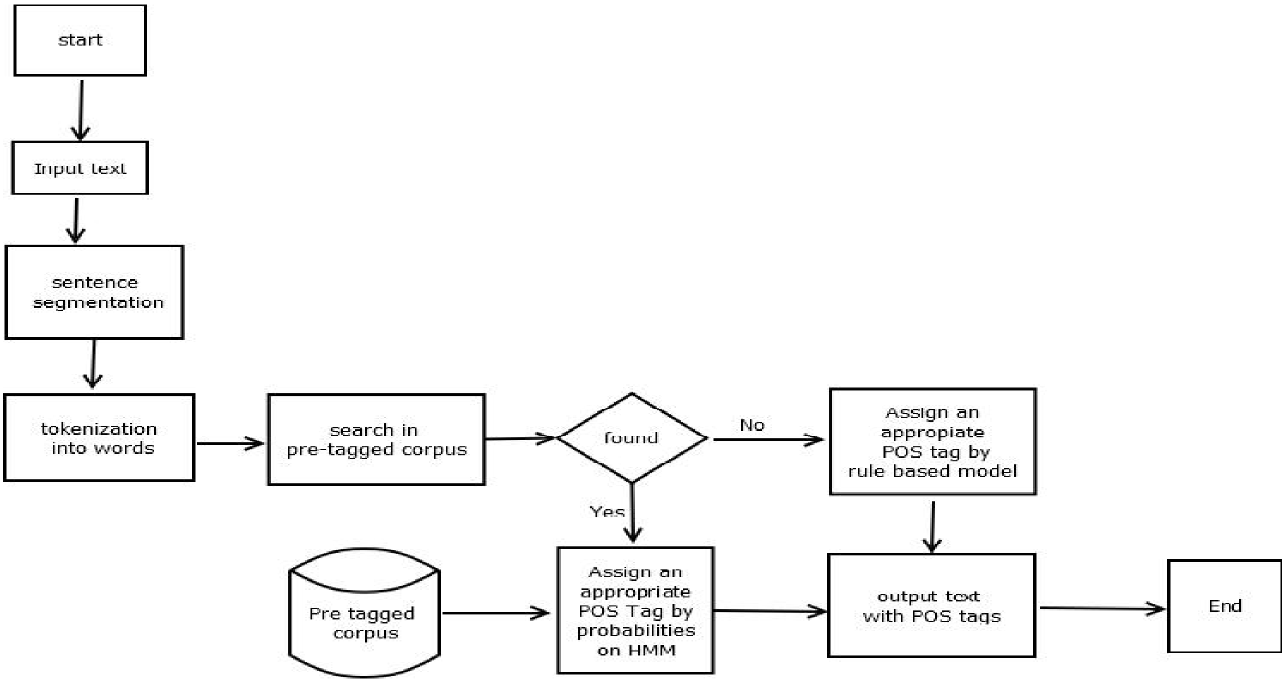
start
Input text
sentence segmentation
tokenization into words
search in pre-tagged corpus
found
No
Yes
Assign an appropiate POS tag by rule based model
Pre tagged corpus
Assign an appropriate POS Tag by probabilities on HMM
output text with POS tags
End

Fig. 1 .A Hybrid Model for POS tagging

A. Input text

The first stage only takes the input data and check the data is in Devanagari Hindi format. The input can be inserted into text area or can open any Hindi text file.

B. Sentence segmentation

Next stage we divide input data into a sentence using "Puranviram"(|) or "prashanvachakchinha"(?). We get the output data in the form of an individual sentence.

C. Tokenize into words

The sentence is now tokenized into words by splitting the words using “space”. The Devanagari Hindi input data will be tokenized into individual words.

D. Tagging Hindi data

At this stage tagging of individual data words is done in following steps

Hidden Markov model

Hidden Markov Models (HMM) have been widely used in various Natural language processing tasks to tag words to their corresponding tags like noun, pronoun, verb, adjective, adverb, etc. For example,

Input text

यह एक

Whose output sequence is

यह _PRP एक _QC _RB _NN _VM । _PUNC

The system shows output text words with their corresponding tags. We use $w_1 \ldots w_n$ to denote the word of input text and $t_1 \ldots t_n$ denotes the tag sequence. We assume that each $w_i$can take any value in a finite set W of words. For example, W might be a set of possible words in Hindi, for example V = {यह ,एक , , , ,। ,….}. Each $t_i$can take any value in a finite set T of possible tags. For example, K might be the set of possible part-of-speech tags for Hindi, e.g. T = {PRP, QC, RB, . . .}.Define S to be the set of all sequence pairs $w_1 . . . w_n$, $t_1 . . . t_n$ such that $n \geq 0$, $w_i \in W$ for i = 1 . . . n and $t_i \in T$ for i = 1 . . . n. This is called the tag sequence or state sequence This type of problem, where the task is to map a sentence $w_1..w_n$ to a tag sequence $t_1..t_n$, is often referred to as a sequence labeling problem, or a tagging problem.

- A parameter q(s|u, v) for trigram (s, u, v) such that $s \in T \cup \{STOP\}$, and u, $v \in W \cup \{*\}$.The value for q(s|u, v) can be resulted as the probability of seeing the tags after the bigram of tags (u, v).
- A parameter e(x|s) for any $x \in W$, $s \in T$. The value for e(x|s) can be resulted as the probability of seeing observation x paired with state s.

$$P(W_1 = w_1 .. W_n = w_n, T_1 = t_1 .. T_{n+1} = t_{n+1}) = \prod_{i=1}^{n=n+1} P(T_i = t_i | T_{i-2} = t_{i-2}, T_{i-1} = t_{i-1}) \prod_{i=1}^{n} P(W_i = w_i | T_i = t_i) \quad (1)$$

$$P(T_i = t_i | T_{i-2} = t_{i-2}, T_{i-1} = t_{i-1}) = q(t_i | t_{i-2}, t_{i-1}) \quad (2)$$

and that for any value of i, for any values of w_i and t_i,

$$P(W_i = w_i | Y_i = t_i) = e(w_i | t_i) \quad (3)$$

$$P(W_1 = w_1 .. W_n = w_n, T_1 = t_1 .. T_{n+1} = t_{n+1}) = \prod_{i=1}^{n=n+1} q(t_i | t_{i-2}, t) \prod_{i=1}^{n} e(w_i | t_i) \quad (4)$$

Rule Based Model

After applying HMM model we now apply regular expressions (build on finite state machines) tag text in form of, time and date, numbers, punctuation marks, special symbols like 11/Mar/11, 14/3/14, 10:09 am, 16:00, 123,56 , etc and assign tags to input text respectively. The regular expressions increase the efficiency of the system.

In the final stage of Part-of-speech tagging system applies various grammar rules depends on the previous and next tagged word or by using a suffix or prefix rules.

E. Output text

The text contains words with their respective taggers in the sequential order and the words which are not tagged are tag by "NONE" in the output text and represented in output text area.

EXPERIMENTS AND RESULTS

A POS tagger system generally gives functionalities as sentences segmentation and tokenization of words and POS tagging for input text. This system also provides these functionalities with 100% correctness of segmentation and tokenization functionalities and 92.56% precise POS tagging functionality. Experiments are done in various domains like news, history, stories, etc and performance is evaluated on these domains.

Experiment 1: Sentence Segmentation:

This experiment will result in splitting of input text in sentences. For example,

Input:

अब इन कम
कर अब

Output:

1.अब इन कम
कर

2. अब

Experiment 2: Tokenization into words

This experiment use spaces between words to tokenize input Hindi text into words. For example,

Input:

अब इन कम
कर

Output:

अब ,इनकम , , , , , , , , , , , , ,क
र , , ,।

Experiment 3: POS-Tagging

This experiment shows system functionality of "Part-of-speech tagging of input Hindi text".

For example,

Input:

अब इन कम
कर

Output:

अब_RBइनकम _NN _NN _NN _NN _VNN _PREP _PREP
_NN _PREP _NN _NN _PREP _RB कर _VFM _VAUX
_VAUX । _PUNC

Evaluation

The system is evaluated on various data sets. A measure of evaluation is precision and accuracy of the system.

Precision

$$= \frac{\text{total no of word tagged correctly}}{\text{Total number of tagged words}} \qquad (5)$$

$$\text{Accuracy} = \frac{\text{t n o w t c}}{\text{T n o w}} \qquad (6)$$

The system using HMM model increase the precision and accuracy by using a statistical approach. It also removes ambiguity by using probabilistic approach. The rule based on the Hindi language also improves the quality of tagging. The system yield 92.56% of average precision and 87.55% of average accuracy.

Conclusion and Future Scope

In this paper, we have presented a Hybrid approach (HMM and rule based model) for Part of Speech Tagger of Hindi Language. This tagger improved performance of Natural Processing System which assigns to a word the most likely tag assigned to that word in the training corpus. So, the present system is still under the development, especially in morphological knowledge acquisition. As the future work, we hope to increase precision and accuracy of our system with increasing grammatical rules in Hindi Language and will classify taggers more precisely which give us deep knowledge of a language without increasing the size of tagged corpus.

References

1. Brill E., 1992. A simple rule-based part of speech tagger, Proceedings of the Third Conference on Applied Natural Language Processing ANLC '92, Stroudsburg, PA, USA, 1992, 152–155,.
2. Zin K. K. and Thein N.L., 2009. Part of speech tagging for Myanmar using hidden markov model, Proc. International Conference on the Current Trends in Information Technology (CTIT), 2009, Dubai, Dec 2009, 1–6.
3. AniketDalal, Kumar Nagaraj, Uma Sawant and SandeepShelke. (2006). Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach, In Proceeding of the NLPAI Machine Learning Competition, 2006.
4. Smriti Singh, Kuhoo Gupta, Manish Shrivastava, and Pushpak Bhattacharyya. (2006). Morphological Richness Offsets Resource Demand-Experiences in Constructing a POS Tagger for Hindi, In Proceedings of Coling/ACL 2006, Sydney, Australia, July, pp.779-786.
5. AgarwalHimashu, AmniAnirudh. (2006). Part of Speech Tagging and Chunking with Conditional Random Fields, In the proceedings of NLPAI Contest, 2006.
6. Garg N., Goyal V and Preet S., 2012. Rule based Hindi part of speech tagger, Proc. Coling, Mumbai, India, December, 2012.